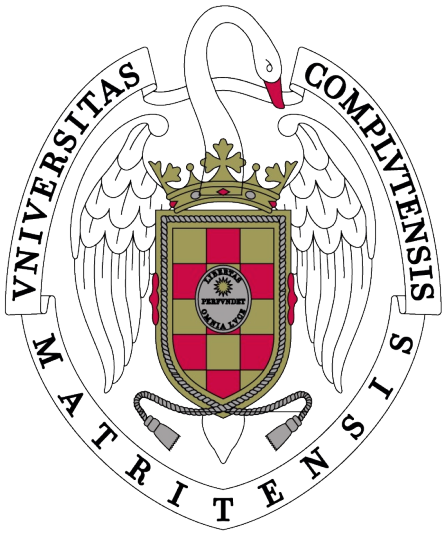

Proyecto Fin de Máster en Sistemas Inteligentes

"Máster en Investigación en Informática, Facultad de Informática, Universidad Complutense de Madrid"

Personalización de perfiles de usuario en entornos móviles

Autor: Alejandro Palacios Provencio

Director: Alberto Díaz Esteban

Curso 2009 / 2010

El/la abajo firmante, matriculado/a en el Máster en Investigación en Informática de la Facultad de Informática, autoriza a la Universidad Complutense de Madrid (UCM) a difundir y utilizar con fines académicos, no comerciales y mencionando expresamente a su autor el presente Trabajo Fin de Máster: “Personalización de perfiles de usuario en entornos móviles”, realizado durante el curso académico 2009-2010 bajo la dirección de Alberto Díaz Esteban en el Departamento de Ingeniería del Software e Inteligencia Artificial, y a la Biblioteca de la UCM a depositarlo en el Archivo Institucional E-Prints Complutense con el objeto de incrementar la difusión, uso e impacto del trabajo en Internet y garantizar su preservación y acceso a largo plazo

AGRADECIMIENTOS

Quiero agradecer a todos mis compañeros y familiares que han participado de alguna manera en este trabajo, tanto de forma logística como la psicológica.

Todos los que han soportado las eternas encuestas para llevar a cabo la evaluación en esta investigación y haberme escuchado cuando les hablaba sobre mis cosas con demasiado entusiasmo. Gracias a mis compañeros de despacho, Elisa, Gabriel y Álvaro, que compartíamos el esfuerzo de los trabajos que estabamos realizando en nuestras tertulias de desayunos y comidas. A Yaofeng, que fue un fichaje de última hora para colaborar con la evaluación y realizó todas las encuestas a pesar de todo el trabajo que tenía. Todos saben que estaré ahí cuando lo necesiten. Gracias a mi famila, a mis padres, Diego y Marisol, a mi hermano, Juan Diego, mis primos, Miki y Marina, y al amor de mi vida, Paola. Otros grandes sufridores de las encuestas de evaluación y que me han apoyado y se han interesado mucho en este trabajo.

Un agradecimiento especial al futuro Doctor en Diseño de Interfaces de Dispositivos Móviles, Álvaro Alcázar Bueno, sin su ayuda, conocimientos y consejo el proyecto carecería de todo estilo y belleza visual.

Por último, pero no menos importante, a mi director de Proyecto, Alberto Díaz Esteban, por confiar en mi y darme tantas oportunidades.

ÍNDICE DE CONTENIDO

ABSTRACT

During last decades, the electronic information, that it has been stored in Internet, has reached a critical point where the users are not able to know what information is valuable for them and which one is not. This problem has been addressed for decades through the adaptability and customization processes.

We have focused the scope of this research Master Thesis on the personalization of the contents over the journalism field, and furthermore, we carry out it over mobile devices what we consider that it is an engaging feature.

On one hand, it is proposed a representation, as generic as possible to cover any kind of information, of each piece of information, which means of each new and user profile.

On the other hand, it is addressed also the personalization over time by browsing and interacting with the system, which interprets the actions of the users to learn their profile in an implicit way.

Using these approaches, it has realized a brief assessment with some ideally constructed users and another largest one, with ten real users divided by the interest groups according to their reflected interest over the complete set of journalistic information.

In each type of evaluation, it has been followed various approaches based on the different parts of the new in question and on the weight of each part.

Finally, the conclusions reached by studying the evaluation results, clarify that the best approach when customizing content, is the approach based on the selection of the header and the summary of a new and give more weight to the first one than the second one.

As future work we propose the use of other terms when representing the units of information and the user profile, such as named entities and the use of concepts. In addition, to make a customization over the position of the user, it is attempted the insertion of a geo-location component.

RESUMEN

A medida que pasa el tiempo, la información electrónica que se tiende a acumular en Internet ha llegado a un punto crítico. Este punto, en el que el individuo ya no es capaz de alcanzar a conocer, de toda esa información, cuál es relevante para él y cuál no. Este problema se agudiza en los dispositivos móviles, donde la información que se puede mostrar es mucho más reducida.

En este proyecto de Máster en Investigación se ha investigado sobre la personalización de contenidos periodísticos en dispositivos móviles.

En primer lugar es necesario disponer de un modelo de usuario que permita reflejar los intereses de cada individuo. Este modelo servirá para seleccionar las noticias que son más interesantes para cada usuario. En particular, en esta investigación se propone una representación, tanto de las noticias como de los perfiles de usuario, formada por términos. Por otro lado, se plantea una propuesta de adaptación a lo largo del tiempo a través de la interacción del usuario con el sistema. Esta adaptación se realiza de manera implícita interpretando las acciones del individuo en su navegación por las noticias que recibe. Todas las acciones tienen un efecto inmediato en la adaptación respecto al contenido y a la presentación, por lo que el perfil del usuario se encuentra en una adaptación continua.

Con todo lo anterior se ha realizado una breve evaluación con usuarios idealmente construidos y otra más extensa con diez usuarios reales, divididos mediante grupos según el interés que reflejen. En cada tipo de evaluación se han tomado varios enfoques basados en las diferentes partes de una noticia (titular, resumen y contenido) y en el peso de las mismas.

Las conclusiones a las que se llega mediante los resultados de la evaluación reflejan que el mejor enfoque a la hora de personalizar contenidos es mediante el enfoque de tomar el titular y el resumen de una noticia, dando más peso al primero que al segundo.

KEY WORDS

Adaptation, personalization, news, user profiles, content text, information retrieval, crawler, recommender systems, vector space model, Rocchio.

PALABRAS CLAVE

Adaptación, personalización, noticias, perfil de usuario, contenido textual, recuperación información, araña web, sistemas de recomendación, modelo espacio vectorial, Rocchio.

I. INTRODUCCIÓN

1 Motivación

La mayoría de sistemas de información suelen tener el problema, y de forma más concreta, Internet, de la gran cantidad de información que van acumulando. El contenido de esta información tiene a menudo una compleja estructura y no sigue ningún patrón definido, estando siempre presente la variabilidad espacial y temporal, a parte de los criterios de credibilidad. Este exceso de información produce sobrecargas hasta tal punto que es probable que el usuario, dependiendo de su nivel de conocimiento, no sea capaz de interpretarla.

Hay que añadir que la forma de presentar todos estos contenidos carece de diseño de interfaces, habiendo una primacía prácticamente absoluta del diseño gráfico sobre la información, con la consecuencia de un desconocimiento de la estructuras de las páginas web, produciendo en el usuario una sensación de desorientación.

Este problema asociado a Internet se agudiza más cuando se trabaja en el dominio de los dispositivos móviles. Estos dispositivos tiene limitaciones en cuanto al tamaño de la pantalla, la interacción del usuario, la velocidad de acceso, etc.

Por tanto, es necesario encontrar una forma de conseguir vencer estas limitaciones. Para ello son especialmente adecuadas las técnicas de personalización de contenidos, que permiten seleccionar para cada usuario aquello que más le interesa.

Fijando el objetivo en la personalización de información, se considera que las tres ideas más importantes son: conocer bien al usuario, adaptar sus intereses a lo largo del tiempo de manera implícita y adaptar los contenidos presentados según sus intereses.

La primera idea es que hay que conocer bien al usuario, a la persona que use el sistema, lo cual hará que el sistema proporcione información de una manera que no se le presentaría a otro usuario.

La segunda idea consiste en no exigir al usuario una realimentación explícita sobre lo que le interesa o no le interesa. La forma de obtener esta información debe estar basada en la interacción implícita del usuario con el sistema.

La tercera idea está relacionada con la forma de presentar la información a los usuarios. Los intereses de lo usuarios deben de reflejar tanto el contenido como la presentación visual del mismo.

2 Personalización de información

La personalización, como se define en varias fuentes, es la adaptación de un producto, servicio o contenido a una persona o usuario, en función de sus características, preferencias personales o información previa que proporciona.

Hay que entender como usuario, cualquier individuo, grupos de individuos, organismos o instituciones (un niño, una mujer, una empresa…), por lo tanto, se puede deducir que personalización significa adaptar algo según unas preferencias.

La personalización hoy en día está presente en todos los ámbitos, desde algo corporativo, como una empresa que tiene bolígrafos con su logotipo, hasta tartas con el escudo o fotografía de alguna institución o individuo, coches que se modifican según los gustos de su propietario, pasando por televisiones que seleccionan solos aquellos programas que más gustan al usuario.

Este trabajo se centra básicamente en la personalización dentro del ámbito del

contenido textual, pero hay que tener presente que la personalización es un termino muy extendido y presente en muchos aspectos cotidianos. La personalización también esta muy presente en todo lo que rodea el mundo del marketing, la publicidad y las ventas. Es muy utilizada en Internet, en las páginas web y también en el marketing online (Amazon,AdSense de Google…) donde se muestran anuncios, productos o contenidos según el perfil del usuario.

El modelado de usuario forma parte del área de la interacción entre hombres y máquinas en la que muchos profesionales desarrollan modelos cognitivos de los usuarios. Se pretende incluir modelos de sus habilidades y conocimiento declarativo.

El resultado o la finalidad que tiene el modelado de usuarios en los sistemas informáticos está en construir y almacenar perfiles de usuario. Un perfil de usuario consiste en una colección de los intereses de información.

3 Fases del trabajo

Las fases que se van a seguir en este trabajo de investigación para alcanzar una solución son las siguientes:

- Primero, si se quiere personalizar una cierta información, es importante qué tipo de información. Por lo que en primer lugar hay que decidir el campo que se quiere personalizar. Además, se ha de recopilar toda la información de una fuente en concreto, o de varias, sobre el tema que se ha elegido. La información debe ser fiable y variada para que el sistema pueda ser capaz de clasificar, adaptar o personalizar cada elemento de información respecto al perfil del usuario
- Segundo, se ha de conseguir una representación del perfil del usuario lo suficientemente sólida como para que para cada elemento de información se personalice de manera correcta.
- Tercero, se ha de encontrar una forma de establecer una similitud entre el perfil de un usuario y cada elemento de información para saber en qué grado le conviene al usuario.
- Cuarto, se tiene que establecer cómo se va a realizar la adaptabilidad del perfil a los contenidos, es decir, la realimentación. Si se va a utilizar algún tipo de factor de olvido o la relevancia que tendrán ciertas partes de la información para ajustar el perfil.

4 Estructura de la memoria del proyecto

La memoria se estructura en cinco partes. La presente introducción, en la que se ha pretendido dar una vista general del propósito del trabajo y poner en contexto al lector identificando el problema a tratar, el objetivo que se quiere conseguir mediante este trabajo y la solución que se propone.

El segundo capítulo consiste en un estado del arte del campo de la personalización de información. Temas como el modelado de usuario y personalización de información son prácticamente obligatorios, además de la adaptación de los contenidos a lo largo del tiempo.

En el tercer capítulo se pasa a la descripción del sistema propuesto. En primer lugar, la descripción y un breve análisis de la arquitectura general que se ha utilizado. Segundo, se concreta qué tipo de información se utilizará para poner en práctica el

sistema de personalización y las fuentes de información convenientes para obtener los elementos de información. Tercero, se barajan varios perfiles de usuario en el sistema concluyendo en una estructura específica. Y para finalizar el capítulo, las metodologías de adaptación y similitud de la información.

El cuarto capítulo muestra los experimentos que se han llevado a cabo mediante las metodologías de personalización y similitud de la información descritas en el capítulo anterior.

Finalizando esta memoria se encuentra el capítulo quinto, que trata las conclusiones a las que se llega mediante los datos obtenidos en el capítulo anterior. Se comparan los resultados de las metodologías empleadas y se proponen algunas ideas a tener en cuenta para continuar este trabajo.

II. Estado del arte

1 Introducción

Antes de desarrollar e implementar el sistema o la metodología de personalización, se han analizado un conjunto de trabajos sobre los campos que se listan a continuación:

- Modelado de usuario, se puede considerar la piedra angular de cualquier sistema de personalización. Es el filtro mayoritario de todo el contenido, pueden existir otros parámetros, pero sin lugar a dudas, una buena representación del perfil de un usuario es el buen camino para obtener unos resultados adecuados.
- Selección de información, también es importante determinar cuales son las técnicas que se utilizan para seleccionar los contenidos que más interesan a cada usuario. Estas técnicas estarán muy relacionada con el modelo del usuario. Por otro lado también esta relacionada con la representación de la información que se va a recomendar. También se pueden tomar como referencia algunos sistemas de adaptabilidad o personalización que se están usando en este momento.
- Adaptación a lo largo del tiempo de los intereses de los usuarios. Existen distintas técnicas para ajustar los perfiles de usuario a lo largo del tiempo, unas se basan en realimentación explícita del usuario y otras en la determinación de forma implícita de esos cambios.
- Adaptabilidad en la presentación de los contenidos en dispositivos móviles. Aparte de personalizar los contenidos textuales, no hay que olvidar su interfaz. Aunque existen unos patrones ya definidos, no siempre se ajustan a cualquier persona en particular.

2 Modelado del usuario

2.1 Introducción

Según (Tu y Hsiang, 2000),a la hora de construir un perfil de un usuario individual, la información puede obtenerse explícitamente, con la intervención directa del usuario, o implícitamente, mediante métodos que monitoricen la actividad del usuario. El grado de modificación de un perfil también hay que tenerlo en cuenta, es decir, se trata de un perfil que puede modificarse o que posee ciertos argumentos, que toma el nombre de *perfil dinámico*, en contraste con el *perfil estático*, que mantiene la misma información desde el momento en el que es creado.

Para que un sistema tenga constancia de la entrada o la presencia de un usuario, la identificación del mismo es importante para saber qué información hay que cargar para la adaptación a dicho usuario. Una vez este dentro del sistema, es vital que se vaya aprendiendo qué contenido es relevante. Para ello existen dos maneras de llevarlo a cabo. Por un lado, la recolección de información del usuario de *forma explícita*, es decir, a menudo pide al usuario que realice algún comentario o bien rellene algún formulario para conocerle mejor y de esta manera clasificar de una forma correcta la información que se le debe de proporcionar, este método es muy usado por sistemas comerciales. La otra

forma de recolección de información sobre el usuario es de *forma implícita*, cuya principal ventaja es que no requiere de intervención adicional del usuario durante la construcción del perfil. Sobre el método de recolección implícito existen varias técnicas que se citan en (Gauch et.al, 2007), basadas en cachés, en proxys, en el uso de agentes (Lee, Chen y Jian, 2003) o mediante logs. Cada una de ellas recoge un tipo de información diferente como la historia o la actividad de navegación, búsquedas frecuentes o bien, todo lo anterior, en el caso de las técnicas basadas en agentes de escritorio, pero cada una de estas técnicas tienen sus pros y sus contras.

Los perfiles de usuario se representan generalmente bien como un conjunto de palabras clave con un cierto peso, redes semánticas o conceptos o asociación de reglas. Los perfiles *basados en palabras clave* son los más simples de construir, porque se encargan de obtener todas las palabras que podrían ser interesantes para contenidos futuros, aunque para que proporcionen un resultado realmente bueno, el usuario debe de utilizar el sistema con regularidad para que sea capaz de construir una buena base de datos con esas palabras clave. Otra representación que pueden tener los perfiles, son los *basados en redes semánticas*, donde están representados por una red semántica con pesos añadidos, en la cual, cada nodo representa una palabra o conjunto de palabras y todo esta relacionado entre sí, con un cierto peso; de la misma manera que la representación anterior, para un buen funcionamiento, es necesario una re-alimentación adecuada para obtener buenos resultados. La última representación de un perfil de un usuario es mediante *conceptos*, que es similar a la representación basada en redes semánticas, es decir, con nodos y relaciones entre ellos, pero en el caso de un perfil basado en conceptos, los nodos representan temas abstractos que se consideran interesantes para el usuario, en lugar de palabras específicas o grupos de palabras relacionadas.

En la figura 3.1 se muestra el flujo de construcción de un perfil de usuario similar al que proponen en (Tu y Hsiang, 2000), pero se ha modificado de una forma más específica a los intereses de lo que se propone en este documento, que consisten en tener una recopilación de información implícita de la cual se pueda construir el perfil del usuario con la representación que mejores resultados obtenga en la etapa de evaluación.

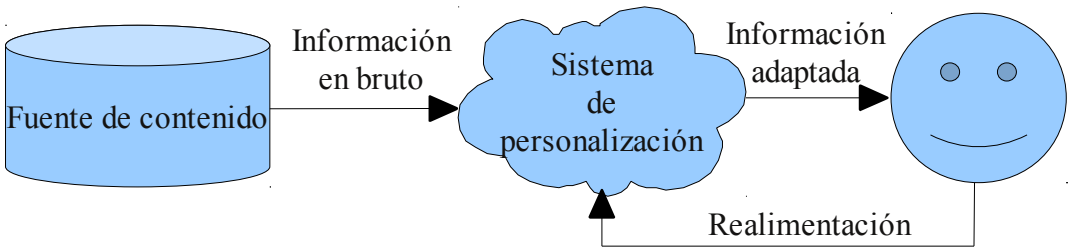

Figura 3.1. Estructura general de un sistema de personalización

2.2 Modelos concretos

2.2.1 Estereotipos

El modelado mediante estereotipos se basa en la identificación de grupos que tengan ciertas características homogéneas respecto a una aplicación (Rich, 1979). Para cada estereotipo se identifican un pequeño número de características clave que permiten al sistema identificar si para el usuario se puede clasificar o no dentro de un determinado grupo. De esta forma a un usuario concreto se le puede aplicar uno o varios estereotipos.

Los estereotipos son muy importantes al principio de una sesión de personalización ya que permiten proporcionar un conocimiento por omisión sobre el usuario que posteriormente se actualizará con los datos obtenidos con la interacción(da Cruz, García y Romero, 2003).

2.2.2 Términos

El método más simple y más habitual de representar al usuario, cuando se trata de establecer sus preferencias con respecto al contenido de los documentos, consiste en que el usuario introduzca un conjunto de palabras clave o términos que representen sus intereses. Adicionalmente se le puede asignar un peso a cada una de esas palabras para determinar con más exactitud cuál es la importancia asociada a cada una de ellas. Estos conjuntos de palabras clave con pesos suelen ser tratados como vectores en el MEV.

2.2.3 Parejas (atributo/valor)

En relación con el modelo anterior, se pretende agrupar mediante una lista de parejas el conocimiento del usuario (Wu et al., 2000). Se entiende por conocimiento, los gustos y preferencias del usuario. Estos atributos pueden ser conceptos o simples palabras,como en (Gauch et al., 2007), que tiene como valor un conjunto de palabras, o bien, un número para indicar su relevancia, como muestra la figura 3.2.

$$\langle clave_{1,} valor_1 \rangle, \langle clave_{2,} valor_2 \rangle, \ldots, \langle clave_n, valor_n \rangle$$

Figura 3.2. Esquema de la representación atributo valor

2.2.4 Categorías

Otra forma de definir los intereses de los usuarios es a través de categorías representadas por conjuntos de términos obtenidos de conjuntos de documentos clasificados previamente en cada una de esas categorías. Esta forma de definir los intereses es similar a tener varios perfiles compuestos por vectores de pesos de términos, que representan distintos temas de interés del usuario. Sin embargo, la diferencia estriba en que los perfiles pueden cambiar mediante la realimentación del usuario, mientras que la representación de las categorías es, en principio, estática. Los usuarios eligen las categorías que mejor representan sus intereses y les pueden asignar un peso para determinar más exactamente la importancia que les otorgan.

La determinación de similitudes entre documentos y categorías es estudiada por un tipo de técnica de clasificación de texto denominada categorización de texto (Sebastiani, 1999). Esta técnica consiste en determinar la asignación de documentos a un conjunto de categorías previamente definidas. Los documentos pueden ser de cualquier tipo de elemento de información con contenido textual: noticias, artículos, páginas web, correos electrónicos, etc. Las categorías pueden tomarse a partir de cualquier sistema de categorías previamente definido: categorías de una biblioteca, categorías de directorios de Internet, etc.

2.2.5 Modelo SEM-HP

En un SEM-HP(modelo sistémico, evolutivo y semántico para el desarrollo de sistemas hipermedia adaptativos), el sistema de aprendizaje soporta la mayor carga de

adaptación (García-Cabrera, 2001). Éste inicializa y mantiene el modelo de usuario y en función de su contenido realiza la adaptación sobre la navegación. El modelo de usuario es actualizado automáticamente mientras que el usuario navega. La información que almacena puede dividirse en dos tipos:

- Características estáticas. Características del usuario que nunca cambian o que cambian con muy poca frecuencia. Estas características no dependen directamente de la navegación.
- Características variables. Cambian frecuentemente durante la navegación.

La tabla 3.1 describe el contenido del modelo de usuario. Las características variables se encuentran en las filas 1–5, las características estáticas en las filas 6-9.

CARACTERÍSTICA	DESCRIPCIÓN
Meta	Información que el usuario desea conocer.
Conocimiento	Valor de conocimiento del usuario sobre cada ítem de la EC de navegación. Es un número entre 0 y 100.
Items leídos	Items leídos por el usuario.
Número de lecturas	Número de veces que el usuario a leído un ítem.
Items relevantes	Items para cuya lectura el usuario esta preparado.
Experiencia en la materia	Conocimiento general del usuario sobre el dominio conceptual y de información del SHA.
Experiencia en el hiperespacio	Pŕactica del usuario en el uso de sistemas hipermedia.
Preferencias	Predilección del usuario por determinados medios, tipos de documentos, etc.
Datos personales	Datos de tipo edad, sexo, profesión, etc. que serán utilizadeos para inicializar el modelo de usuario usando estereotipos.

Tabla 3.1. Contenido del modelo del usuario SEM-HP

3 Selección de información

3.1 Introducción

La personalización de información se puede considerar también como un tipo de filtrado o adaptación de contenidos. Dentro de este campo, aparecen varias cuestiones, como por ejemplo, qué se entiende por filtrado o personalización de contenidos, cómo se diferencia esta adaptación de una simple recuperación de información, porqué se necesita, qué significa la personalización para una persona, cómo se usaría y con qué propósito.

Las características básicas de la personalización consiste en tener un conjunto dinámico de documentos, información estable y especificada, un perfil o modelo de usuario y un proceso de selección. Con todo esto, el filtrado se puede definir como el proceso de determinar que perfiles tienen una alta probabilidad de cumplir un objetivo en particular dada una fuente de contenido, por otro lado, la recuperación son simples

consultas a base de datos o almacenamiento de información.

La necesidad de la adaptación de información está más que justificada por varios motivos. Primero, el crecimiento de Internet es exponencial. Segundo, uno de los impactos de esta red de redes es que cualquier persona con acceso a ella puede convertirse en autor y publicar contenido, como consecuencia, la cantidad de información a encontrar es extremadamente diversa y la cantidad de información disponible es enorme, lo que conlleva a una sobrecarga de información. Por otro lado, la información que es muy importante para un individuo no tiene porqué serlo para otros, como comenta Bowman (Bowman et al., 1994) “alrededor del 99% de la datos disponibles no son interesantes para alrededor del 99% de los usuarios”. En consecuencia se necesitan sistemas que obvien la información irrelevante basada en las preferencias del usuario, aquí es donde entran en juego los perfiles de usuario que se explica más detalladamente en la siguiente sección de este capítulo.

En esta sección se analizan los procedimientos básicos en un sistema de adaptación o personalización y, seguidamente, se hace un repaso a algunos modelos de adaptación, como el que se utiliza en el caso de estudio de “PlaSerEs” o el SE(modelo sistemático, evolutivo y semántico para el desarrollo de sistemas hipermedia adaptativos).

3.2 Identificación del usuario

Con el fin de poder adaptar el contenido al usuario hace falta, antes que nada, identificarlo para poderlo diferenciar del resto. Después hay que mostrar aquella información que desea o que es interesante para él. Un usuario se puede identificar de una manera directa o indirecta.

3.2.1 *Identificación Directa*

Es aquella en que el usuario se identifica él mismo de forma personal y directa. Sería el caso en que introduce su nombre y clave (cuando es necesario), de esta manera el sistema lo reconoce de forma inequívoca. También sería el caso que selecciona uno de los perfiles disponibles, como por ejemplo en el sistema operativo de un ordenador donde hay varias sesiones creadas.

3.2.2 Identificación Indirecta

Es aquella en que la identificación se produce de una manera transparente al usuario, donde él no realiza la acción de identificarse, sino que hay un procedimiento paralelo que lo hace. Sería el caso, por ejemplo, de los navegadores que tienen guardadas los datos del usuario en las cookies, o cuando se introduce una tarjeta de crédito en un cajero y este nos identifica por el número de la tarjeta. La identificación indirecta es menos eficaz que la directa.

3.3 Métodos de selección de contenido

Una vez el usuario esta identificado y se saben sus preferencias, es necesaria una descripción del contenido para poder diferenciar aquel que cumple los requisitos solicitados del que no los cumple. La descripción de contenidos se realiza mediante indexación.

Sintetizando los contenidos de los artículos (Sarawagi, 2008) y (Billsus y Pazzani, 2007), que tratan la adaptación de información de una posible fuente de contenidos. A la

hora de acceder a contenidos adaptados, en este caso, se pueden tener en cuenta cuatro tipos diferentes de adaptabilidad, como se menciona en (Billsus y Pazzani, 2007), que se centran en mayor parte en la personalización de noticias periodísticas, pero para este artículo se han intentado generalizar. La primera forma de adaptabilidad consiste en la *personalización de contenidos*, que se trata de añadir un ranking de contenidos y comparar este ranking con el modelo de usuario. Este tipo de adaptabilidad se centra en técnicas de adaptación que modelan los intereses de los usuarios basados en sus comentarios explícitos o implícitos, y usa el resultante modelo de usuario para personalizar el contenido de la información.

La segunda forma de adaptabilidad es mediante *navegación*, cuyo objetivo es simplificar el acceso a contenido relevante. Esta técnica se centra en analizar los patrones de acceso del usuario para determinar la posición de los elementos de menú dentro de la jerarquía del propio menú. El sistema estima la probabilidad de que un usuario seleccione una determinada opción que esté incluida en un menú específico y se usa esa probabilidad para la construcción de menús los cuales son los que que tienen más probabilidades de contener opciones que el usuario seleccionará. Una reordenación sencilla de un menú con adaptabilidad es un punto fuerte, ya que no requiere de una infraestructura compleja dependiente del contenido basado en perfiles de usuarios individuales, lo que significa que es mucho más fácil de implementar y satisfacer los requisitos de escalabilidad en el mundo real frente a técnicas más complejas.

La tercera forma de adaptabilidad consiste en el *acceso a contenidos contextuales*, que a veces se denomina recuperación en tiempo real, y está estrechamente relacionada con la personalización basada en contenido. Sin embargo, en lugar de utilizar el modelo de los intereses del usuario aprendido con el tiempo, el enfoque se basa en que la información que se muestra actualmente, ya sea una página web o un mensaje de correo electrónico, éstos se muestren como una expresión de los intereses actuales de los usuarios, es decir, a muy corto plazo.

Por último, otra forma de adaptabilidad consiste en la *agregación de contenidos*, que son servicios que automáticamente agregan información desde diferentes fuentes de contenidos y, como resultado, adapta el conjunto de elementos de información más actuales. Un servicio como este puede ser implementado usando una ponderación de términos estadísticos y técnicas de similitud de textos que evalúan la similitud entre dos tipos de contenidos, además, los enfoques de la categorización de textos puede usarse para entrenar clasificadores que automáticamente fijen una categoría a nuevos contenidos de diferentes proveedores en un conjunto de nuevas categorías.

Para aprender el modelo de usuario se pueden utilizar los *intereses a corto plazo*, cuyo propósito es doble, primero, contiene la información más recientes, así que otro tipo de información que pertenezca al mismo hilo de eventos pueden ser identificadas, y segundo, permite identificar información que el usuario ya sabe; una opción para archivar estos términos es con el algoritmo de los vecinos más cercanos (Pazzani y Billsus, 2007). Por otro lado, los *intereses a largo plazo*, intentan modelar las preferencias generales de un usuario, ya que la mayoría de las palabras que aparecen en las noticias no son útiles para este fin, el sistema periódicamente selecciona un vocabulario adecuado para cada nueva categoría individual.

Actualmente la personalización es una práctica muy utilizada y cada vez más presente, ya que se implementa en casi todos los sistemas o espacios donde hay contenidos. Con la llegada de la TDT y de los centros multimedia es muy útil poder acceder a los contenidos audiovisuales que uno desea y una de las funcionalidades que

permiten muchos de los dispositivos es la grabación de programas o búsquedas de video y es aquí donde aparece la personalización para poder, por ejemplo, grabar aquellos programas que cumplen nuestras preferencias de una forma automática, GoogleTV.

3.3.1 Selección basada en el MEV

La forma más habitual de realizar la selección de contenidos cuando se utilizan términos con pesos en el MEV es mediante la aplicación de la fórmula del coseno (ecuación (2.1)) entre los vectores de pesos de términos que representan a los documentos y el vector de pesos de términos que representa al modelo de usuario. Los documentos clasificados con mayor similitud (por encima de un umbral o un número fijo de ellos) son los que son seleccionados para el usuario.

Cuando el modelo de usuario almacena varios vectores de términos se pueden combinar los resultados obtenidos para cada uno de los vectores, habitualmente seleccionando la máxima similitud vector-documento. También se puede dar mayor peso a los términos que aparecen en el título de los documentos frente a los que aparecen en el cuerpo (Nakasima y Nakamura, 1997).

$$MEV(P,N)=\frac{\sum P_i \cdot N_i}{\sqrt{\sum P_i^2}\cdot\sqrt{\sum N_i^2}}$$

Ecuación 3.1. Fórmula del coseno del modelo de espacio vectorial

3.3.2 Selección basada en el modelo bayesiano

Una de las técnicas más populares utilizadas para clasificar son los algoritmos bayesianos ingenuos (Henze y Nejdl, 1999), que aplican los conceptos de la probabilidad de Bayes dentro del problema de clasificación, para obtener un clasificador bayesiano formado por un único nodo raíz C, que representa la clase variable y un nodo Xi por cada una de las características. Así, dado el vector de características x = (x1, …, xn) de un documento d, la probabilidad de que d pertenezca a la categoría viene determinado por la ecuación 3.2.

$$P(C=c|X=x)=\frac{P(X=x|C=c)\,P(C=c)}{\sum P(X=x|C=c)\cdot P(C=c)}$$

Ecuación 3.2. Probabilidad de que d pertenezca a una categoría

En la ecuación anterior, la probabilidad P(X=x | C=c) es difícil de estimar sin imponer restricciones de simplicidad, puesto que los posibles valores de X son demasiados si se tiene en cuenta que los términos están relacionados o son dependientes. La solución a este inconveniente consiste en simplificar la red bayesiana asumiendo que los atributos X1, …, Xn son condicionalmente independientes dada la categoría C, (de ahí surge el nombre de clasificador ingenuo). Esto permite rescribir la ecuación del siguiente modo:

$$P(C=c|X=x)=\frac{P(C=c)\cdot\prod P(X_i=x_i|C=c)}{\sum P(C=c')\cdot\prod P(X_i=x_i|C=c')}$$

Figura 3.4. Ecuación final de la probabilidad entre un documento y una categoría

Resultando ahora más sencilla la estimación de los términos P(X | C) y P(C) a partir de las frecuencias del corpus de entrenamiento. Pese a su simplicidad, la efectividad de un clasificador Naïve Bayesian está avalada por una gran cantidad de estudios empíricos a pesar de la hipótesis de independencia.

3.3.3 Selección basada en el modelo probabilístico y episódico

Este modelo se basa en que, dado un documento y una pregunta, es posible calcular la probabilidad de que ese documento sea relevante para esa pregunta. Discutir la equiparación probabilística requiere algunas explicaciones sobre teoría de probabilidad, por lo que se asume que en un momento dado podemos utilizar cualquier respuesta a una pregunta. Así, todas las probabilidades discutidas se toman en el contexto de esa pregunta. Si se asume ésto, para el propósito de la discusión, el número de documentos de la base de datos que son relevantes a la pregunta son conocidos. Si un documento es seleccionado aleatoriamente de la base de datos hay cierta probabilidad de que sea relevante a la pregunta. Si una base de datos contiene N documentos, n de ellos son relevantes, entonces la probabilidad se estima en:

$$P(rel)=\frac{n}{N}$$

En concordancia con la teoría de la probabilidad, la de que un documento no sea relevante a una pregunta dada viene expresada por la siguiente fórmula,

$$P(rel_{inv})=\frac{N-n}{N}$$

Obviamente, los documentos no son elegidos aleatoriamente, sino que se eligen sobre la base de la equiparación con la pregunta —basado en el análisis de los términos contenidos en ambos—. Así, la idea de relevancia está relacionada con los términos de la pregunta que aparecen en el documento.

Una pregunta dada divide la colección de documentos en dos conjuntos: los que responden a la pregunta y los que no. Sin embargo, todos los documentos seleccionados no son realmente relevantes. Entonces, debemos considerar la posibilidad de que un documento sea relevante o no, dado que haya sido ya seleccionado. Si se supone que un conjunto de documentos S de la base de datos ha sido seleccionado en respuesta a una pregunta. La cuestión es hasta qué punto éste es el conjunto que debería haber sido seleccionado en respuesta a la pregunta. Un criterio debe ser seleccionar el conjunto si es más probable que un documento del conjunto sea más relevante que otro que no lo es.

Evidentemente, la recuperación probabilística envuelve muchos cálculos y premisas. Numerosos experimentos demuestran que los procedimientos de recuperación probabilística obtienen buenos resultados. De cualquier forma, los resultados no son mucho mejores que los obtenidos con un modelo booleano o vectorial. Posiblemente en el nuevo contexto de la recuperación a texto completo de bases de datos heterogéneas en Internet, compliquen lo suficiente la recuperación como para que las técnicas de recuperación probabilística se utilicen más. Un ejemplo de ello es el trabajo de Gövert,

Lalmas y Fuhr que utilizan el enfoque probabilístico para facilitar la categorización de documentos en la web (Gövert et al., 1999).

3.3.4 Selección basada en el vecino más cercano

El algoritmo del vecino más próximo (Nearest Neighbour, NN) es uno de los más sencillos de implementar. La idea básica es como sigue: si calculamos la similitud entre el documento a clasificar y cada uno de los documentos de entrenamiento, a qué de éstos mas parecido nos estará indicando a qué clase o categoría debemos asignar el documento que deseamos clasificar.

Desde un punto de vista más práctico, el vecino más próximo puede aplicarse con cualquier programa de recuperación de tipo best macth. Lo más frecuente es utilizar alguno basado en el modelo vectorial, pero, en puridad, esto no es imprescindible. Lo necesario es que sea best macth y no de comparación exacta, como pueda ser el caso de los booleanos; el algoritmo de basa en localizar el documento más similar o parecido al que se desea clasificar. Para esto no hay más que utilizar ese documento como si fuera una consulta sobre la colección de entrenamiento.

Una vez localizado el documento de entrenamiento más similar, dado que éstos han sido previamente categorizados manualmente, sabemos a qué categoría pertenece y, por ende, a qué categoría debemos asignar el documento que estamos clasificando. Una de las variantes más conocidas de este algoritmo es la del k-nearest neighbour o KNN que consiste en tomar los k documentos más parecidos, en lugar de sólo el primero. Como en esos k documentos los habrá de varias categorías, se suman los coeficientes de los de cada una de ellas. La que más puntos acumule, será la candidata idónea.

4 Adaptación a lo largo del tiempo

4.1 Adaptación explícita frente a adaptación implícita

La adaptación explícita del usuario tiene la característica de que mucha de la información sobre el usuario es añadida mediante acciones especificadas en alguna parte del sistema por los propios usuarios. Los usuarios preguntan directamente al sistema, pero no siempre saben como realizar la pregunta para que el sistema les responda adecuadamente, un ejemplo de ello son las consultas en algún buscador, como en Google, que puede ocurrir no obtener la información adecuada porque no se tiene el conocimiento necesario para realizar la pregunta.

La adaptación implícita del usuario es construida por el sistema a través de la interacción del usuario. Pose un comportamiento más sofisticado ya que se han de calcular suposiciones, razonar sobre las creencias de un usuario, inferir los planes del usuario, etc... El sistema necesita estar seguro de los nuevos hechos y puede ser necesario que resuelva conflictos.

4.2 Adaptación basada en el algoritmo de Rocchio

Cuando se utilizan términos representados en el MEV la adaptación se realiza mediante el ajuste de los pesos de los términos o la adicción de nuevos términos extraídos de los documentos indicados como relevantes, o no relevantes, por los usuarios.

El algoritmo de Rocchio (Rocchio, 1971) es uno de los algoritmos de aprendizaje

más utilizado para clasificación de texto. La adaptación para un usuario, (*Qo* consulta original, Qm, consulta resultado), se recalcula para incluir un porcentaje arbitrario de los documentos relevantes (*Dr*) y no relevantes (*Drn*), como una forma de enseñar al motor de búsqueda, y, posiblemente, aumentar la precisión. El número de documentos relevantes y no relevantes permitidos que se dan a la entrada está definidos por los por los pesos a, b, c. La formula de Rocchio es la que sigue:

$$\vec{Q}_m = (a \cdot \vec{Q}_o) + (b \cdot \frac{1}{|Dr|} \cdot \sum \vec{Dr}_j) - (c \cdot \frac{1}{|Dnr|} \cdot \sum \vec{Dnr}_k)$$

Ecuación 4.1. Fórmula original de Rocchio

4.3 Adaptación basada en el algoritmo bayesiano

Las probabilidades del clasificador bayesiano ingenuo se pueden ir ajustando con los nuevos documentos juzgados por los usuarios. Las probabilidades de las redes bayesianas se pueden ir ajustando con los nuevos documentos juzgados por los usuarios de manera similar a como ocurre con el clasificador bayesiano ingenuo.

4.4 Adaptación basada en el vecino más cercano

La adaptación basada en el vecino más cercano se produce cuando se introducen documentos relevantes. El cálculo del vecino más cercano a la llegada de un nuevo documento será ahora distinto por la presencia de más documentos relevantes.

5 Adaptabilidad de la presentación en dispositivos móviles

5.1 Introducción

La experiencia de usuario es el conjunto de factores y elementos relativos a la interacción del usuario, con un entorno o dispositivo concretos, cuyo resultado es la generación de una percepción positiva o negativa de dicho servicio, producto o dispositivo (Deagostini y Cormenzana, 2005). Depende no sólo de los factores relativos al diseño (hardware, software, usabilidad, diseño de interacción, accesibilidad, diseño gráfico y visual, calidad de los contenidos, buscabilidad o encontrabilidad, utilidad, etc) sino además de aspectos relativos a las emociones, sentimientos, fiabilidad del producto, etc.

Además de esta diversificación y complejidad, se asiste a una descentralización del acceso a la información, delegando al usuario el trabajo de buscar lo que quiere conocer. Hay que encontrar una manera de mejorar el aprovechamiento de los sistemas informáticos para hacerlos más sencillos de usar y aprender, a parte de lograr que la interacción sea efectiva e intuitiva.

5.2 Definiciones

Para que en este trabajo se tenga una buena experiencia de usuario es importante una buena adaptatividad, que no hay que confundir con adaptabilidad. Esta segunda consiste en permitir al usuario modificar los parámetros del sistema para adaptarlo a su comportamiento, mientras que la primera, se entiende como la capacidad del sistema de adaptarse automáticamente al usuario, basado en suposiciones del mismo.

Un sistema adaptativo, según (Benyon, 1994), es “aquel que, basado en el conocimiento, altera automáticamente aspectos de funcionalidad e interacción para lograr

acomodar las distintas preferencias y requerimientos de sus distintos usuarios." Como ejemplos de comportamiento adaptativo se pueden citar la presentación de formularios y menúes dependiendo de la tarea a realizar, la presentación de información relevante según la tarea o usuario que la demande, o el ofrecimiento de ayuda según el contexto de trabajo.

Pero en un dispositivo móvil no se puede abrumar al usuario con formularios o encuestas de adaptatibilidad, por lo que habría que encontrar alguna forma de obtener la misma información que se hace de forma explícita con los formularios y encuestas, de una forma implícita.

Los objetivos a marcar para esta investigación se pueden deducir de (Deagostini y Cormenzana, 2005), donde se observan determinadas características:

- Mejorar la eficacia y eficiencia de los sistemas informáticos
- Extender el rango de usuarios, desde el novato al experto
- Satisfacer las demandas del usuario, reduciendo temores y aumentando el atractivo y la flexibilidad, logrando así una mejor aceptación
- Incrementar la productividad
- Reducir la curva de aprendizaje
- Exceso de información
- Permitir el diálogo entre el usuario y el sistema
- Presentar información de manera integrada y comprensible

5.3 Modelado de usuario para la interfaz gráfica

De la misma manera que para el buen funcionamiento de un sistema personalización de información es importante la presentación del usuario con un modelo lo suficientemente preciso, también es la clave para los buenos resultados adaptativos en los interfaces que comunican a la máquina con el humano. Los usuarios suelen diferir en:

- Cómo debe usarse un ordenador
- El modelo mental del sistema que está usando
- Sus habilidades físicas y perceptivas
- Sus capacidades delo que conoce
- Sus destrezas para identificar y realizar tareas

Por otro lado, si se toma como referencia a los autores (Deagostini y Cormenzana, 2005), éstos identifican las diferencias de los usuarios como las siguientes:

- Habilidades psico-motoras
- Competencia
- Capacidad de aprendizaje
- Comprensión de problemas y tareas asociadas
- Expectativas
- Motivación
- Preferencias

El modelo de usuario permite adaptación en distintas tareas, alcanzar diferentes objetivos, adecuarse a necesidades cambiantes y atender distintos niveles de conocimiento y destreza. Algunos ejemplos con buenos resultados usando modelos de usuarios son sistemas de filtrado de información, adaptación de la presentación de resultados, adecuación de la interacción y sistemas inteligentes de enseñanza.

Para construir un modelo de usuario respecto al entorno gráfico, como en este

caso, puede aparecer la aparición de ruido o cambios en el usuario referente a sus gustos y hay que tener en cuenta el detalle y la flexibilidad que se le proporcionará al modelo.

El papel del modelo del usuario respecto a las interfaces gráficas se puede interpretar como las acciones del usuario, según sus opciones y el historial del diálogo con el sistema, que luego generarán respuestas del sistema, tanto a nivel lógico como físico.

5.4 Adaptatividad en interfaces gráficas

Para adaptar el contenido visible se puede realizar de dos formas, automática y a petición del usuario.

La adaptatividad automática puede dividirse en cognitiva y operativa. La cognitiva trata de emplear los métodos mediante los cuales el ser humano procesa la información. Si bien éstos son relativamente estables y cambian con lentitud, es difícil medir las diferencias entre los mecanismos empleados por distintos individuos, lo que complica el desarrollo de este tipo de interfaces. Una posible solución al problema consistiría en identificar los componentes de tareas complejas para determinar luego las habilidades cognitivas necesarias para realizarlas.

La *operativa* consiste en detectar y analizar el comportamiento del usuario para predecir acciones futuras y adaptar la interfaz para facilitarlas. Es sin duda la opción más empleada y que ha mostrado mejores resultados, partiendo del supuesto que las acciones del usuario identifican generalmente el objetivo a alcanzar, permitiendo así adecuar la respuesta del sistema y predecir con éxito las intenciones y preferencias del usuario.

La *adaptitividad colaborativa* implica que el usuario define cuánto y qué puede ser adaptable en un sistema. Por ejemplo, las especificaciones del usuario pueden ser directas, como cambiar el sistema contextual de la ayuda, o puede ser un fino cambio de nivel en los tipos de intervenciones del sistema: demora antes de la identificación de ciertos objetos, sugerencias de acciones, etc. Otra posibilidad es darle la oportunidad de definir sus objetivos. El usuario podría elegir de un conjunto de objetivos generales y el sistema podría sugerirle diferentes formas específicas para lograr el objetivo señalado, por ejemplo: explorar, planificar, buscar, evaluar su propio progreso, etc. Luego seguirá interactuando de acuerdo al modo que haya seleccionado dependiendo de dónde y cómo quiera poner su atención y qué tipos de actividades prefiera realizar.

Es indiscutible la importancia que tiene la interfaz Humano Computador, en el sentido que puede afectar directamente el rendimiento de un sistema y la productividad de los usuarios. El concepto clave en el funcionamiento de este tipo de interfaces es el de “modelo de usuario”. A través del mismo se pretende sintetizar las características y habilidades de un grupo de personas con el fin de facilitar y mejorar la interacción con el sistema. La adaptatividad del mismo se logra intepretando las acciones del usuario, según sus opciones y el historial del diálogo con el sistema, y generando respuestas tanto a nivel lógico como físico. El proceso no es sencillo: no existe el “usuario promedio”, el conocimiento del usuario no es estático y es prácticamente imposible crear modelos precisos. Pero se puede desarrollar un modelo razonablemente válido que permita mejorar la *adaptatividad* de este tipo de interfaces.

6 Ejemplos de sistemas de personalización

6.1 ASNA

Los autores (Al Masun et al., 2006) admiten que el análisis de opiniones favorables o desfavorables o de afinidad de emociones es una tarea que requiere la inteligencia emocional y un profundo conocimiento del contexto textual, con la participación de sentido común y conocimiento del dominio, así como los conocimientos lingüísticos.

La interpretación de las opiniones es generalmente un asunto discutible para los seres humanos. Sin embargo, el sistema, ASNA, es un intento para llevar a cabo tarea. El enfoque de este sistema es sencillo.

En primer lugar, el usuario elige las fuentes de las noticias de acuerdo a su ámbito de interés. En este caso, utilizan RSS que te toman como la fuentes de contenido. Después de que las fuentes de contenidos se seleccionen, la obtención de contenidos recopila las noticias como tuplas de tema de noticias y la historia breve que corresponde al tema, al analizar los resultados devueltos por los canales RSS. A continuación, las tuplas de texto son analizados por un analizador de lenguaje. Han implementado una técnica de análisis del sujeto, Asignatura Tipo, Asunto atributos; Acción, Posición de acción, atributos de acción y de objetos, Tipo de objeto y atributos de los objetos para cada línea de texto. La salida del analizador de lenguaje se evalúa mediante una herramienta lingüística SenseNet que han desarrollado utilizando WordNet (Fellbaum, 1999) y ConceptNet (Liu y Singh, 2004). SenseNet considera cada tupla como un sentido y genera un valor numérico para cada unidad léxica (una frase, por ejemplo). A continuación, el motor de emociones clasifica las noticias en función de ocho tipos de emociones que son, feliz, triste, esperanzado, temeroso, admirable, vergonzoso, amable y odio, más una categoría neutra. Por último, un usuario puede navegar por las noticias de acuerdo a los grupos de emoción.

6.2 PIA-System

Un sistema PIA (P*ersonal Information Agent*), que según sus autores (Albayrak et al., 2005), conoce la forma en que piensa un usuario y puede ayudarle durante todo el día mediante el acceso, filtrado y presentación a cualquier tipo de información.

El sistema de agentes de PIA esta compuesto por varias clases de agentes. En primer lugar, agentes extractores de información, por un lado son motores de búsqueda general, pero, por otro, los portales del dominio específico han de estar integrados en el sistema. Segundo, agentes que implementan distintas estrategias de filtrado que han de estar combinados de una manera inteligente. Tercero, agentes para proveer a diferentes tipos de presentaciones de interfaces y un agente personal para cada usuario, que obtiene el perfil del mismo mediante una forma implícita.

En el desarrollo y evaluación de su sistema PIA, comentan que la información es obtenida mediante WWW, e-mail, SMS, MMS y clientes J2ME, donde el sistema adapta la presentación, usando las preferencias del perfil con un tamaño apropiado para un teléfono y una PDA. La recopilación la realizan de manera constante los agentes de extracción de información, que una vez insertada una nueva información alertan al agente gestor del modelado. Para una recuperación eficiente de la información se realiza un pre-procesado, que primeramente consiste en usar distintas tablas en la base de datos global para diferentes dominios, seguidamente, se construyen varios modelos de cada documento y se indexan para una óptima recuperación. El filtrado de información se

puede hacer dependiendo del domino de la información, por lo que habrá varias implementaciones en diferentes ámbitos (noticias, documentos científicos, conferencias...). Por último, en el nivel de la presentación, el sistema provee de diferentes métodos de acceso adaptando la interfaz al dispositivo del usuario.

En este sistema con PIA, la parte de mezclar la lógica del contenido de datos y la parte de presentación no me ha parecido apropiada, como a cierto grupo de desarrolladores que pueden llegar a ser demasiado puristas a la hora de la construcción de una aplicación, aunque hay que tener en cuenta que el hecho de saber cómo se han de presentar los datos requiere un cierto nivel de lógica de datos y no sólo de estética.

6.3 Recomendaciones de programas de televisión

Toman en cuenta dos tipos de enfoques a la hora de desarrollar su sistema de recomendación. Por un lado, un sistema *personalizado*, es decir, basado en contenido y del que se ha estado hablando todo el documento. Pero existe otro enfoque, el *colaborativo*, que consiste en recomendar elementos a un usuario específico de acuerdo con la evaluación de otros usuarios con gustos similares. Los autores de (Lee y Yang, 2003) proponen un híbrido que combine ambos enfoques.

Por otra parte,como se describe en su artículo, disponen de tres tipos de agentes que se utilizan para la recopilación de información, el modelado del usuario y la gestión de tareas que se incluyen en el sistema de adaptación. Los agentes que se dedican a la recopilación de información, obtienen dos tipos de programas de televisión, películas y noticias, por lo tanto usan un *agente de películas* y otro *agente de noticias*. El agente de películas trabaja de dos maneras, puede proporcionar un formulario en el que el usuario tiene que, explícitamente, señalar sus puntos de interés, o bien, de una forma implícita, el propio sistema registra y analiza los contenidos que el usuario ha leído. Para los programas de noticias, el agente que se encarga de ello, extrae todo el contenido del diario on-line de la CNN y organiza la información antes de presentarla, para que el usuario, de una forma explícita, indique sus intereses.

El agente para el modelado del usuario se encarga de construir el modelo de las preferencias del usuario desde la información registrada en su perfil. Los agentes, para modelar al usuario, en realidad, se subdividen en dos tipos, el *agente de aprendizaje*, que tiene la responsabilidad de construir un sistema de clasificación que consiste en un conjunto de modelos de predicción, y el *agente de colaboración*, que está implementado en un enfoque de decisión el cual es uno de los métodos más populares y puestos en práctica para una inferencia inductiva. El modelo de aprendizaje está representado como un árbol de decisión, que, una vez se tiene una instancia de algún tipo de contenido, la clasifica recorriéndole desde la raíz a las hojas llegando a especificar la categoría de dicha instancia, los autores han definido únicamente dos categorías, *user-like* (gusta al usuario) y *user-dislike* (no gusta al usuario).

Por último, en el *agente dedicado a la adaptabilidad* esta diseñado con el propósito de clasificar las películas en sus clases apropiadas, dependiendo del resultado del análisis. Para clasificar las noticias, su sistema tiene en cuenta todos los canales de noticias como la misma clase. De manera que sólo determinar si se recomienda un cierto canal de noticias a un usuario. El orden de los diferentes canales de noticias está predefinido por el usuario. Por lo tanto, para los programas de noticias, el procedimiento de realimentación del aprendizaje se activa manualmente. El usuario puede añadir nuevos ejemplos arbitrariamente o eliminar los viejos del perfil para obtener un nuevo modelo.

La solución para la recomendación de programas de televisión que han propuesto los autores desde luego me parece completamente viable, no sólo, por la organización, que es sencilla, pero al mismo tiempo robusta, si no porque, en términos prácticos, es un sistema que llegaría a la gran mayoría de la gente por ser la televisión unos de los principales medios de entretenimiento. Para terminar el análisis de este sistema, propongo que una componente de geo-localización haría que el sistema se adaptara aun más al propio usuario, ya que no tienen en cuenta los contenidos en canales de televisiones regionales, que muy frecuentemente tienen contenidos que pueden ser de interés para el usuario, como meteorología, accidentes u otros tipos de eventos.

6.4 MyPortal

Otro sistema que apuesta por utilizar agentes inteligentes es el que se propone en (Tu, Mine y Amamiya, 2009), haciendo uso de la arquitectura de computación P2P. La arquitectura conceptual de su Web semántica se construye basándose en cuatro idas principales.

Primera, es usar una arquitectura de computación peer-to-peer con énfasis en un método eficiente para reducir la carga de la comunicación, ya que un sistema centralizado tiene cuellos de botella y su coste de mantenimiento es elevado, en cambio, los escalables y descentralizados sistemas P2P están recibiendo cada vez más especial atención en los campos de la investigación y el desarrollo de productos para la web al se un entorno abierto y dinámico. La segunda idea consiste en que todos los participantes contribuyan a las descripciones semánticas de una forma consistente; se proponen tres tipos de participantes, el consumidor que busca fuentes por la Web; el proveedor, que posee ciertos recursos, y el mediador, que habilita la comunicación entre el consumidor y el proveedor. La tercera idea trata sobre la integración de contenidos web con servicios web, ya que los servicios Web proveerán una nueva forma de proporcionar información Web, por lo tanto, una gestión integrada y unificada de contenidos y servicio web necesita ser llevada a cabo a través de diferentes niveles. Y la cuarta idea proporciona un enrutamiento a toda la información de la que los usuarios pueden estar interesados, que suele ser principalmente de dos tipos, *información local*, que es la que esta almacenada en el PC y puede ser accedida sin conexión de red, e *información remota*, que es gestionada y almacenada en un servidor remoto, publicada por el administrador de un sitio web y sólo se puede acceder a ella mediante internet.

En su arquitectura, todos los proveedores y consumidores están construidos como un “MyPortal”, que consiste en una semántica personalizada para satisfacer los requisitos de información de un usuario, proporcionando un camino a la información relevante para el usuario. Cada proveedor describe sus capacidades en lo que ellos llaman WSCD (*Web site capabilitity description*) y cada consumidor enviará consultas relevantes basadas en los requisitos del usuario cuando una búsqueda web sea necesaria. El mediador esta compuesto de agentes asignados al consumidor y a los proveedores, donde ambos pertenecen a una comunidad basada en agentes que se comunican mediante el método P2P de recuperación de información, para cumplir con el intercambio de información entre varios MyPortals.

6.5 PlaSerEs

La plataforma PlaSerEs (Marín et al., 2008) tiene como objetivo principal proveer información de los productos o servicios ofrecidos por establecimientos comerciales a los

clientes de una manera personalizada.

Tienen en cuenta los siguientes términos:

- Dispositivos móviles (DM), que es la ventana donde los usuarios realizarán todas sus gestiones mediante las conexiones inalámbricas.
- Fuentes de información (FI), a las que se conectan los usuarios para obtener contenidos que siempre tienen en cuenta sus preferencias o necesidades.
- Perfiles de usuario y/o modelado de usuarios, lo consideran como un conjunto de criterios para adaptar los contenidos procedentes de las FI.
- Sistema de información (SI), es el sistema que engloba todo lo anterior y se encarga de evolucionar el perfil de usuario durante una sesión.

Para la adaptabilidad en DM, utilizan el estándar de W3C llamado CC/PP (acrónimo de Composite Capabilities/Preferences Profiles) que define las características básicas de un dispositivo de acceso, pero de una manera extendida como realizó (Indulska et al, 2003). que muestra características del usuario, de su sesión, de su DM y de su localización.

Su modelo de adaptación consta de tres módulos que se explican a continuación, a parte de plasmarlos mediante la figura 4.1.

- Módulo de presentación, tiene las características a tomar en cuenta para desplegar los contenidos en el DM. Está compuesto por el modelo de perfil de SM y lo definen usando las extensiones de CC/PP propuestas por Indulska (Indulska , 2003).
- Módulo de contenidos, dentro de este módulo aparece el modelo del perfil de usuario, con las preferencias de gustos e intereses entre otros, y el modelo de contexto, que contempla el clima, localización, hora del día, día de la semana, fecha, días festivos y servicios, entre otros.
- Módulo de conexión inalámbrica. Lo dividen en cuatro módulos. Un módulo de hardware, que toma en cuenta las interfaces de comunicación del DM y de las FI. Un módulo de software, que contempla los protocolos de comunicación y los sistemas operativos soportados en DM y FI. Un módulo lógico, que cuenta con un árbol de decisión, permitiendo seleccionar la tecnología más indicada. Y finalmente, un módulo clasificador taxonómico, que toma varias características de los módulos anteriores con el fin de seleccionar la mejor configuración a ser utilizada por la aplicación.

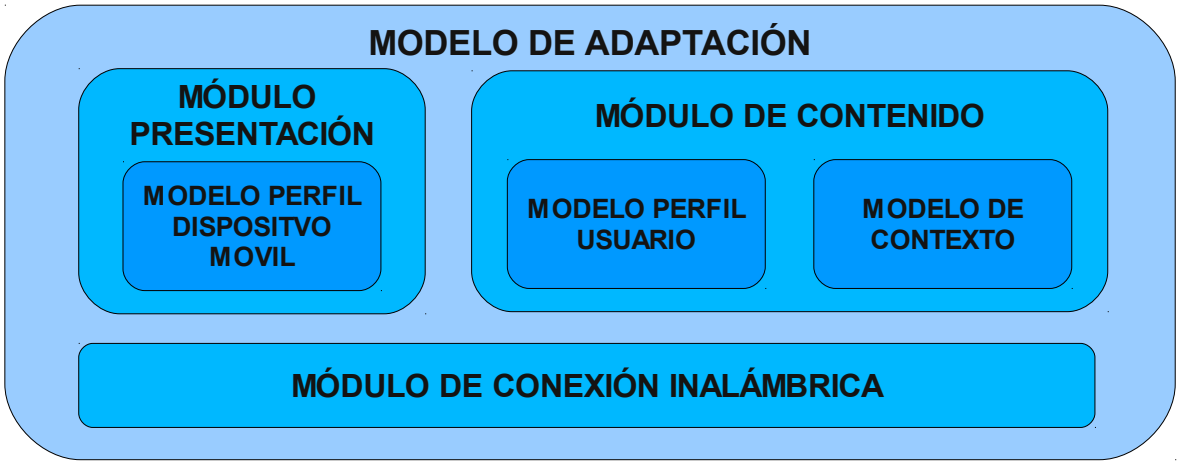

Figura 4.1. Componentes del Modelo de Adaptación de PlaSerEs

Después de explicar el modelo de adaptabilidad que utilizan, pretenden ofrecer los servicios generales a cualquier establecimiento haciendo posible realizar reservas, consultas de productos, información detallada, pedidos, control de cuentas y envío de promociones y mensajes con información general a cada usuario de acuerdo a sus perfiles y lo que más les pueda interesar.

Un caso de uso está realizando en un restaurante de comida típica en la ciudad de Bogotá D.C, Colombia, que, sin demasiado esfuerzo, se ha podido adaptar llegando a la conclusión de que se puede extender a cualquier tipo de establecimiento comercial. Aunque dejan como trabajo futuro el estudio detallado de los componentes de usuario, contexto y dispositivos de acceso.

6.6 SEM-HP

Los SHA (Sistemas Hipermedia Adaptativos) aparecen con el objetivo de mejorar la usabilidad de los sistemas hipermedia tradicionales. La mayoría de ellos consiguen facilitar la actividad del usuario, haciendo que el sistema se ajuste a determinadas características de éste. A la hora de crear un sistema hipermedia hay que plantearse cuestiones como cuál es la funcionalidad del sistema susceptible de adaptación, a las características de quién o qué se ajusta el sistema, qué técnicas y métodos utiliza el sistema para producir la adaptación y en qué momento durante el funcionamiento del sistema se produce la adaptación. Es decir, las preguntas más frecuentes que se hacen los investigadores y desarrolladores a la hora de trabajar con sistemas de adaptabilidad.

Los autores de este modelo siguen la estructura de De Bra (De Bra et al, 2000) en el que existen tres elementos que están presentes en la mayoría de ellos SHA:

- Modelo de dominio, describe la estructura del dominio de la aplicación en términos de conceptos y relaciones entre conceptos
- Modelo de usuario, almacena las características del usuario que el sistema tiene en cuenta para realizar la adaptación. Suele incluir el conocimiento del usuario sobre los conceptos del modelo de dominio.
- Modelo de adaptación, establece cómo la información del modelo de usuario influye en la adaptación del sistema. También especifica cómo y cuando actualizar la información almacenada en el modelo de usuario.

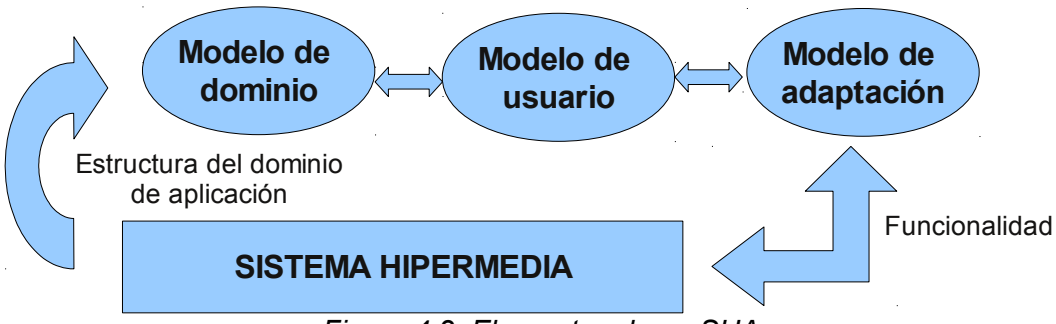

Figura 4.2. Elementos de un SHA

Como exponen en su trabajo un SEM-HP es un modelo sistémico, evolutivo y semántico para el desarrollo de sistemas hipermedia adaptativos. Sistémico porque concibe el sistema hipermedia como un conjunto de sistemas interrelacionados y en interacción. Semántico porque ofrece una semántica flexible con la que el autor puede caracterizar sus dominios de información. Evolutivo porque permite al autor realizar

cambios estructurales y funcionales de una forma fácil y consistente. Adaptativo porque se ajusta a las características concretas de cada usuario. La semántica es una característica muy importante, ya que cuanto más explícita sea la semántica, mayores son las posibilidades de adaptación y evolución.

La arquitectura que proponen está compuesta por cuatro módulos relacionados entre ellos, de los cuales, los tres primeros están explicados con más detenimiento en (García-Cabrera, 2001) y (Medina-Medina et al., 2001):

- Módulo de memorización, almacena, estructura y mantiene el conocimiento. El elemento principal es la estructura conceptual (EC), que consiste en una red semántica con conceptos e ítems. Los enlaces de la EC son relaciones entre conceptos o relaciones entre ítems y conceptos. El SEM-HP proporciona acciones evolutivas para crear, modificar o borrar conceptos ítems o relaciones.
- Módulo de presentación, mediante el filtrado de las EC permite seleccionar un conjunto de los conceptos, ítems y relaciones incluidos en la EC inicial. La evolución de este subsistema hace que se muestren u oculten elementos de la estructura conceptual.
- Módulo de Navegación, permite añadir restricciones que establecen un orden parcial entre los trozos de información ofrecidos por el sistema, es decir entre los ítems de la EC. La evolución del sistema permite la gestión de la misma.
- Módulo de Apredizaje, se encarga de realizar la adaptación del sistema hipermedia. Identifican cuatro elementos principales, modelo de usuario, reglas de conocimiento, reglas de actualización y técnicas de adaptación.

La adaptación del usuario se realiza de manera implícita, es decir, durante la navegación del usuario, por lo que el sistema almacena información de dos tipos, características estáticas, que es información que no cambia o cambia con una frecuencia muy baja, y características dinámicas, que cambian frecuentemente durante la navegación. Este sistema de adaptabilidad comprende las siguientes técnicas de adaptación:

- Anotación de enlaces, cada ítem de la EC visitado con anterioridad por el usuario es anotado indicando el número de visitas del usuario a dicho ítem (anotación textual) y dibujando su borde en color violeta (anotación visual). Además tanto los ítems como los conceptos de la EC son anotados indicando el nivel de conocimiento que tiene el usuario sobre ellos (anotación textual), así el usuario es consciente del conocimiento que tiene en cada momento y de cómo dicho conocimiento aumenta a medida que navega por el sistema.
- Ocultación y deshabilitación de enlaces, las reglas de conocimiento determinan qué ítems puede visitar el usuario y cuales no.
- Estructura conceptual personalizada, en la fase de presentación el autor crea distintas presentaciones. Posteriormente, se selecciona una u otra, dependiendo de las características y preferencias del usuario actual. Es decir, para cada usuario el sistema elige la presentación que mejor se ajusta a sus necesidades.

Como aparece en (Medina-Medina et al., 2001), los sistemas creados presentan una capacidad de evolución completa, incluyendo la adaptación al usuario típica de los sistemas hipermedia adaptativos actuales. A pesar de que la historia de navegación se representa de manera implícita, los sistemas hipermedia diseñados usando el modelo SEM-HP no son dinámicos, ya que los ítems de información son establecidos a priori.

6.7 AIS

Hay investigadores que se inspiran en métodos naturales, como es el AIS, que en biología se define como el sistema de células especializadas y órganos que protegen un organismo de influencias externas. AIS es un tipo de simulación del sistema inmunológico humano donde los antígenos que pudieran atacar a nuestro cuerpo estimulan al sistema inmunológico a producir anticuerpos.

Los autores (Sobecki y Szczepanski, 2007), que realizan una Wiki de noticias inspirándose en este tipo de métodos, de una forma general, identifican claramente las características del modelo de usuario, y se preocupan por el contenido y la representación y utilización dentro de los posibles sistemas, por lo que dividen el contenido principalmente en los datos que se utilizan y datos de usuario que ayudan a identificarle inequívocamente. Pero también identifican problemas a la hora de modelizar al usuario, en generar su perfil, al dar una inicialización apropiada a un perfil recién creado, en el uso de ciertas técnicas de aprendizaje del perfil, identificar comentarios que puedan aportar información relevante, al uso métodos de filtrado de información y a la hora de utilizar técnicas de coincidencias y adaptación del perfil.

Cuando se utiliza un sistema de recomendación, los autores consideran de gran importancia el uso de métodos de filtrado de varias índoles. Primero, un *filtrado demográfico*, que se basa en la información almacenada en el perfil del usuario (conteniendo distintas características demográficas, estos datos sobre el usuario contienen elementos como direcciones, datos de registro, características del usuario y algunos que otros datos de clasificación del cliente) y usa un estereotipado de razonamiento en las recomendaciones. Segundo, un *filtrado basado en el contenido*, que es un método de recomendación aplicado en muchos interfaces de agentes, que toma descripciones del contenido de elementos evaluados previamente para aprender la relación entre un solo usuario y la descripción de los elementos de la noticia. Y tercero, un *filtrado colaborativo*, que se encarga de realizar predicciones automáticas sobre elementos recomendados por una colección y uso de información sobre gustos de otros usuarios(colaboración); los sistemas de recomendación basados en el filtrado de colaboración usan actualmente una matriz de anotaciones, identificando usuarios similares y recomendando elementos altamente puntuados por otros usuarios.

A la hora de llevar a la práctica el uso de un sistema basado en métodos AIS, se utiliza como modelo para la explicación, exploración y predicción de un sistema biológico inmune, o como abstracción de algún proceso inmunológico. A pesar de que los métodos de AIS son relativamente recientes, hay diferentes aplicaciones que ya están en funcionamiento, como la detección de cambios que se usa para combatir los virus informáticos. Aún así, el método de uso de un sistema AIS consta de cuatro pasos en un sistema de recomendación. Primero, el sistema almacena algunas preferencias de la gente en la base de datos. Segundo, cada usuario introduce sus preferencias para algunos temas, como pueden ser películas, y pide recomendaciones en algunos temas que no ha visto con anterioridad. Tercero, el sistema AIS selecciona un grupo de personas, que toman el papel de anticuerpos, quienes tienen similares preferencias con el usuario particular, que toma el papel de antígeno. Y cuarto, la media ponderada de las preferencias para el grupo de personas se calcula por el filtrado de contenido al generar recomendaciones que pide el usuario.

Para controlar el modelo AIS se utilizan unas determinadas ecuaciones en las que hay que señalar que para una recomendación de artículos son muy efectivas, según la evaluación realizada por los autores, en cambio, es menos eficaz para la recomendación

de interfaces de usuario debido a la sencillez de la estructuración de este tipo de información y que, desde el punto de vista colaborativo, no aporta datos relevantes.

III. Sistema de personalización de perfiles de usuario

1 Introducción

El sistema que se propone un sistema..... En el ámbito del perfil de usuario, se opta por una construcción de manera totalmente implícita y con una representación lo más sencilla posible. Pero marcando el objetivo de que sea potente a la hora de que el perfil se use en una función de similitud para clasificar si una noticia gustará o no al usuario. Por otro lado, la forma de obtener información se llevara a cabo por un *web spider* (Chan, 2008), que recorrerá la fuente de contenidos, obtendrá de ella toda la información formateando las noticias en sus tres componentes, el titular, el resumen y el contenido de la noticia. Seguidamente, se llevara a cabo la indexación de las noticias, facilitando el cruce con el perfil del usuario.

Una vez se tiene esto definido, la arquitectura del sistema debe adaptarse para soportar la aparición de dispositivos móviles (teléfonos, PDA, etc...), por lo que el proceso de cómputo y la comunicación debe ser lo menos pesado posible. En consecuencia, se utiliza la arquitectura cliente-servidor, en la que la mayor parte del procesamiento textual la realiza el servidor, mientras que el dispositivo móvil recoge la información implícita para ir construyendo el perfil del usuario.

2 Arquitectura del sistema

El sistema se basa en una arquitectura cliente/servidor. Los clientes, en este caso son los dispositivos móviles que se conectan al servidor que contiene toda la información, tanto de representación de interfaz, como de los datos del perfil del usuario y la información de la fuente de contenidos, además reside la araña web que se encarga de obtener dicha información.

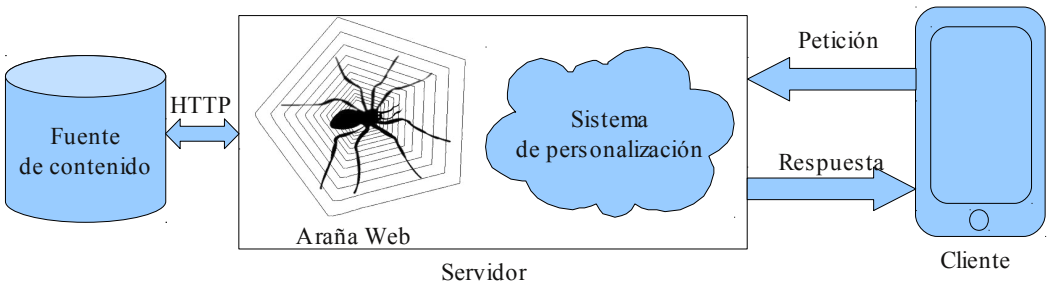

Figura 2.1. Arquitectura del sistema a nivel de comunicación

Este sistema, que se representa en la figura 2.1, aparece una fuente de contenidos donde la información ya se encuentra clasificada en categorías, por lo tanto, a cada uno de los documentos, ya se les asigna de manera trivial su categoría. Cada documento debe de pasar por tres procesos, primero, por un análisis de las diferentes partes del documento, segundo, una indexación de cada una de estas partes y en tercer lugar, una similitud con el perfil del usuario, que alimenta el sistema. Una vez pasado esto, se le ofrece al usuario el contenido que se considera de interés para él.

En una última etapa, se encuentra la navegación por los contenidos de interés o por el resto de documentos que se han obtenido de la fuente por parte del usuario. Aquí se encuentra la obtención de manera implícita, de las preferencias del usuario, que le proporcionan a su perfil la información necesaria para su refinamiento y, de nuevo,

posterior utilización en el sistema de personalización.

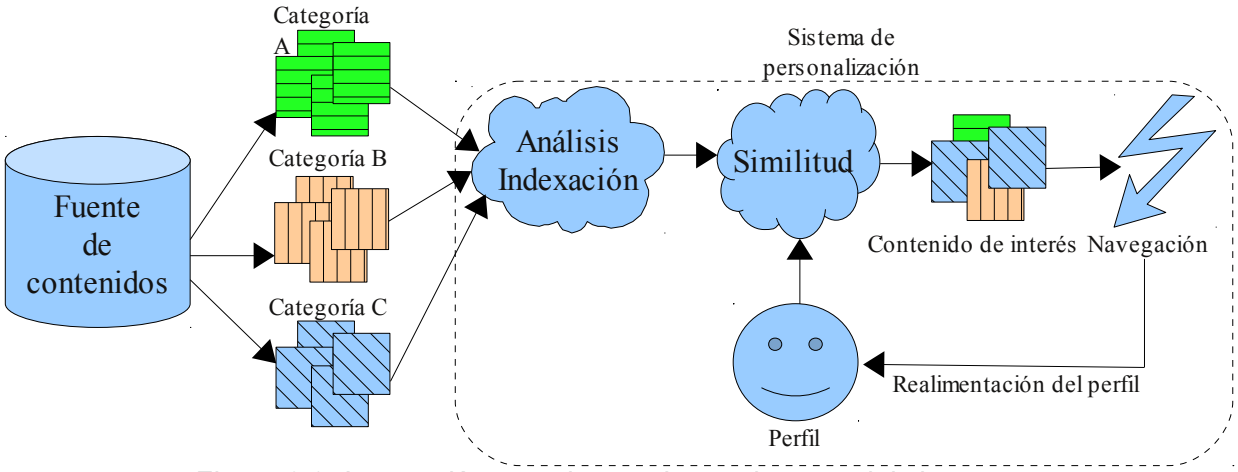

Figura 2.2. Interacción completa en la arquitectura del sistema

3 Obtención de los contenidos

Si se retoman las fases establecidas en el capítulo primero de introducción, en este apartado se resuelven los dos primeros, el primero, que es el tipo de información que se va a personalizar y las fuentes de contenido de las que se obtendrá la información en bruto. Para estos objetivos se va a utilizar técnicas de obtención de contenidos o de información.

Esta tarea dentro del sistema tiene el objetivo de recorrer toda la fuente de contenidos de una forma periódica y automatizada mediante un programa de web spider (araña web). La operación normal es que se le da al programa un grupo de direcciones iniciales, descarga estas direcciones, analiza las páginas y busca enlaces a páginas nuevas. Luego descarga estas páginas nuevas, analiza sus enlaces, y así sucesivamente.

Una descripción breve del proceso que realiza la web spider del sistema consiste en utilizar como direcciones iniciales las direcciones de las diferentes categorías de la fuente de contenido (Chan, 2008) . En cada una de las categorías va recorriendo enlace a enlace, obteniendo las noticias completas, es decir, un titular, un breve resumen, si está disponible y todo el contenido textual y multimedia, si se dispusiera de este último. Un Ejemplos de estas arañas web son Googlebot (Google Webmasters, 2010), Methanol (Methanol, 2010) y Yahoo Web Crawler (Yahoo! Help, 2010).

Pero antes de ponerla a funcionar hay que conocer el dominio que va a recorrer, es decir, las características de la fuente de contenido, el tipo de información que puede y debe obtener, con que tipo de limitaciones se puede encontrar y, por último, pero no menos importante, la frecuencia con que tiene que hacerlo para tener la última información de forma coherente.

3.1 Tipo de contenido

La información en bruto que se ha decidido personalizar han sido documentos de género periodístico. Dentro de este género, existen dos tipos, el de opinión y el de información, habiendo elegido el segundo tipo, es decir, documentos de información, definiendo información como todo aquel texto periodístico que transmite datos y hechos concretos, ya sean nuevos o conocidos con anterioridad. La información no incluye opiniones personales del periodista, ni juicios de valor.

Dentro de este tipo de documentos, son informaciones las noticias, el reportaje informativo y las entrevistas. Sobre las primeras se va a tratar en profundidad ya que son las más demandadas y se utilizarán para el desarrollo de la investigación de este trabajo.

La noticia es todo aquel hecho novedoso que resulte de interés para los lectores. O dicho de otro modo, una noticia es un acontecimiento sorprendente, estremecedor, transcendental y, sobre todo, reciente. La estructura de la noticia consiste en el titular, el resumen y el cuerpo de la noticia que desarrolla la información con todo tipo de elementos complementarios.

Además, una noticia siempre pertenece a una categoría o sección determinada, de forma que el lector sea capaz de encontrar los temas que más llamen su interés.

3.2 Fuente de contenido

La obtención de la información en bruto se realiza desde Europa Press, que genera más de 3.000 noticias diarias, que transmite a través de sus diferentes servicios de texto en un flujo ininterrumpido las 24 horas del día, todos los días del año.

Los servicios de noticias que ofrecen son los siguientes:

- Servicio Nacional: Recoge todos los aspectos de la actualidad en España durante las 24 horas del día con una producción en torno a 500 noticias diarias. El servicio se subdivide en Política, Economía, Sociedad, Cultura, Educación, Laboral, Consumo, Medio Ambiente, Sanidad, Tribunales, Comunicación, Tecnología, Sucesos, con la posibilidad de configurar secciones más específicas aún en función de las necesidades de cada cliente.
- Servicios Autonómicos: Servicios informativos propios ein: Andalucía, Aragón, Canarias, Cantabria, Castilla-La Mancha, Castilla y León, Cataluña, Extremadura, Galicia, La Rioja, Madrid, Murcia, Navarra, País Vasco y Comunidad Valenciana.
- Servicios en Lenguas Propias: Servicios en todas las lenguas del Estado: catalá, valencia, euskera y galego.
- Servicio Internacional: Servicio de información general que recoge los acontecimientos más destacados del mundo. Consta de un promedio diario de 180 noticias y diez crónicas que recogen la información general, política, económica y social.

Optar por esta fuente de contenidos no ha sido por azar, si no porque gracias a ella se podrán alcanzar objetivos como una personalización de grano fino debido a su división en categorías ya pre-establecida. Por otro lado, si se desea tener en cuenta en un futuro el aspecto de la geo-localización en la península y fuera de ella, ésto se cubre con sus servicios Autonómicos e Internacional, contextualizando las noticias según la posición de usuario. También da la posibilidad de personalizar el idioma, proporcionando idas de posibles trabajos futuros.

3.3 Frecuencia de obtención

La fuente no tiene un momento de actualización único, es decir, el contenido se cambia o amplía de forma irregular y existe un momento exacto ni tampoco un patrón de tiempo determinado, por lo tanto, el cambio de la fuente de información es de manera continua y puede ser un problema.

El problema surge debido a que, si tenemos dos usuarios A y B, donde el usuario A solicita información en un momento T_0 y el usuario B solicita la información en un

momento siguiente T_1, relativamente próximo, considerando próximo un intervalo de tiempo entre uno y quince minutos, no reciban la misma información, porque el usuario A no ha llegado a la actualización que se ha producido en el intervalo de tiempo entre la conexión del usuario A y B. Esta situación se debe evitar si se desea un comportamiento coherente y consistente del sistema.

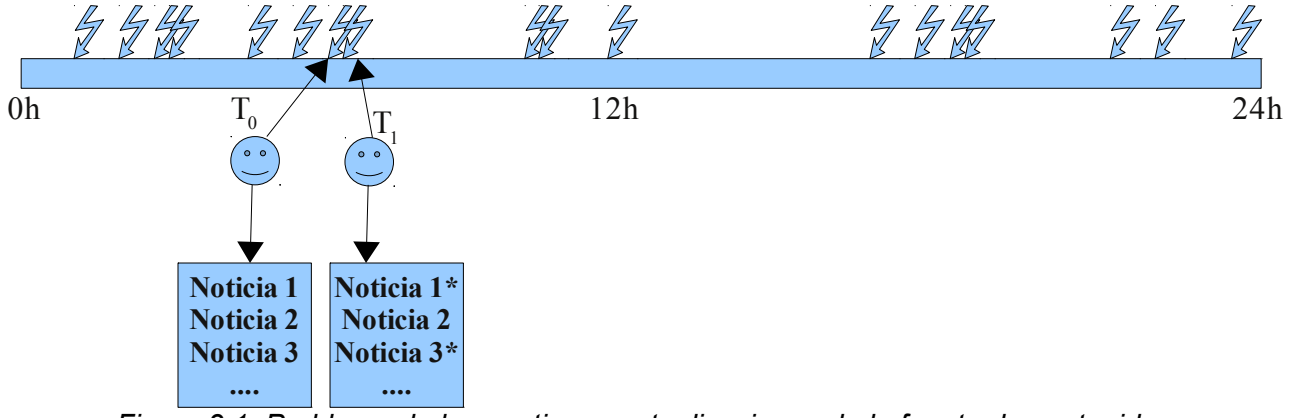

Figura 3.1. Problema de las continuas actualizaciones de la fuente de contenidos

La solución a lo anteriormente expuesto es la discretización de la frecuencia de extracción del contenido. Para ello, se han establecido tres actualizaciones del contenido a unas horas determinadas.

- Actualización matinal, a las seis de la mañana, normalmente, a partir de esta hora existe una mayor afluencia de consultas a las noticias debido a que la gente empieza a levantarse y desayuna mirando lo que ocurre en el mundo.
- Actualización mediodía, a las dos de la tarde, después de la comida mucha gente consulta las noticias, sobre todo en las sobremesas o antes de empezar la segunda parte de la jornada.
- Actualización nocturna, a las diez de la noche, antes de irse a dormir, muchas personas leen un poco las noticias y lo que ha pasado durante el dia.

Estas horas coinciden aproximadamente con las ediciones de diarios de noticias televisivos y de radio, además, en el caso de la actualización matinal, con la entrega de la prensa a los kioskos.

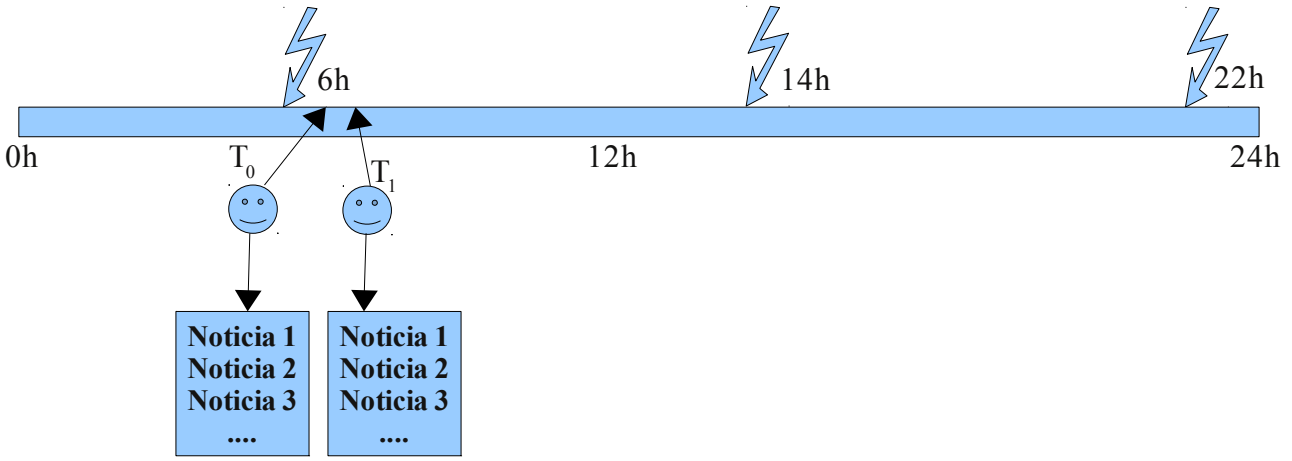

Figura 3.2. Solución propuesta para la discretización de las actualizaciones de los contenidos

3.4 Modelo del contenido

El contenido de la unidad de información se modela como una estructura de datos con tantas tablas como bloques se quiera dividir. Cada tabla consta de dos columnas, una con los términos que aparecen el bloque y otra con la frecuencia de aparición en ese bloque. La representación de forma genérica aparece en la figura 3.3.

Bloque 1		**...**	**Bloque n**	
***Término*$_{bloque1}$**	***Frecuencia*$_{bloque1}$**	...	***Término*$_{bloquen}$**	***Frecuencia*$_{bloquen}$**
Término$_1$	F_1		Término$_1$	F_1
Término$_2$	F_2		Término$_2$	F_2
...	...		...	
Término$_k$	F_k		Término$_j$	F_k

Figura 3.3. Representación general del contenido

Un ejemplo de una noticia de la categoría de ciencias obtenida de la fuente de contenidos EuropaPress aparece en la figura 3.4.

Bloque - Titular		**Bloque - Resumen**		**Bloque - Contenido**	
***Término*$_{tit}$**	***Frecuencia*$_{tit}$**	***Término*$_{res}$**	***Frecuencia*$_{res}$**	***Término*$_{con}$**	***Frecuencia*$_{con}$**
delfines	1	comunidad	1	nariz	1
recurren	1	científica	1	impulsos	3
diplomacia	1	pensaba	1	delfines	2
comunicarse	1	silbidos	1	ráfagas	3
		principales	1	sonidos	5
		sonidos	3	social	1

Figura 3.4. Ejemplo de representación de una noticia

Esta noticia se ha dividido en tres bloques, que consisten en el titular, el resumen y el contenido, y los términos en este ejemplo son palabras. En cada uno aparece la tabla con los términos y su frecuencia de aparición. Como ser verá más a continuación, esta estructura facilita el cálculo de similitud con el perfil del usuario.

4 Modelo del perfil del usuario

4.1 Perfil de categorías

El planteamiento que se tuvo en un principio fue de modelar el perfil de los usuarios únicamente por las categorías a las que se accedía. Es una representación sencilla y poco costosa a nivel computacional, ya que se tenía en cuenta sólo la probabilidad de acceso a cada categoría teniendo una tabla clave-valor.

Categoría	Peso
$Categoría_1$	$Prob_1$
$Categoría_2$	$Prob_2$
…..	…..
$Categoría_n$	$Prob_n$

Tabla 4.1. Representación del perfil de categorías

El principal problema que presenta este modelo de perfil es que las categorías son demasiado generales. Cada una engloba demasiados temas. Por ejemplo, en el caso de la categoría de “Deportes”, dentro de la misma se encuentran el fútbol, el tenis, las fórmula 1, el ciclismo, etc... De manera que, si un usuario esta interesando en el deporte, pero sólo le interesa la fórmula 1, el sistema le devolvería contenidos que para nada son de su agrado por no tener más información.

4.2 Perfil Término-Valor

Como se ha comentado anteriormente, con la información que proporciona la probabilidad de acceso a las categorías no es suficiente, por lo tanto, se planteó la posibilidad de crear una perfil con más información que es de lo que carece lo anterior.

Esta representación consiste también una tabla Término-Valor, de manera que se mantiene su sencillez de representación, aunque su cómputo podría ser algo mas costoso , dependiendo del tamaño de la tabla, ya que en este caso la clave consiste en los términos de las noticias a las que el usuario va accediendo.

Un término se puede interpretar como una palabra en sí, su raíz, se lema o el concepto que representa. El valor es el peso que ese término tiene ha tenido a lo largo de la realimentación.

De esta manera se logra poder diferenciar de alguna manera un tema de otro, aunque el tamaño podría crecer de manera considerable.

Clave	Peso
$Término_1$	$Peso_1$
$Término_2$	$Peso_2$
$Término_3$	$Peso_3$
…..	…..
$Término_n$	$Peso_n$

Tabla 4.2. Representación del perfil por Término-Valor

Sin embargo, aparece otro problema, la ambigüedad de las palabras y sus usos en cualquier categoría. Un ejemplo muy simple es, si se usan palabras como términos, un conjunto de palabras como el siguiente, {“enfrentamiento”, “Argentina”, “Alemania”}, puede aparecer en la categoría “Deportes”, refiriéndose al mundial de fútbol o de cualquier otro deporte. Y por otro lado el mismo conjunto de términos en la categoría “Noticias Internacionales” se puede referir a un conflicto de algún tipo entre ambos países.

4.3 Perfil mixto

Para solventar los problemas de las representaciones anteriores, por un lado la falta de información y por otro, la representación ambigua de un conjunto de términos, se han combinado ambas tablas como muestra la figura 4.1.

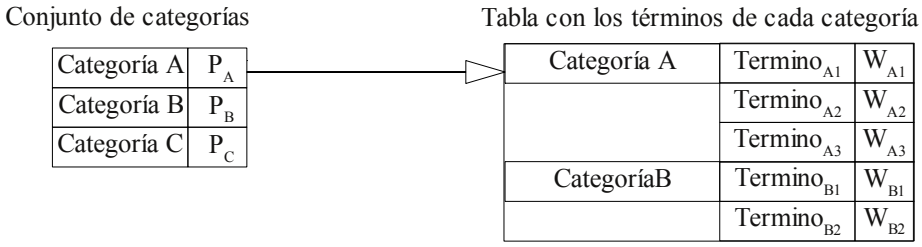

Figura 4.1. Representación del perfil del usuario

Esta representación consiste en un conjunto de categorías con sus probabilidades y una tabla donde se relaciona cada categoría con una cola de prioridad que contiene los términos y sus pesos en esa categoría en concreto. De esta manera, cada conjunto de términos esta relacionado con una categoría y es capaz de definir un tema concreto de la misma.

La computación y almacenamiento de esta representación se puede volver un poco más costosa que las anteriores, ya que, cada categoría tiene un conjunto de términos, y pueden haber repeticiones del mismo conjunto de términos, aunque hay que tener en cuenta que lo más probable es que posean pesos diferentes, con la consecuencia de que no sea viable la idea de una refactorización de alguna manera.

5 Personalización de los contenidos

5.1 Selección de los contenidos

La información que se obtiene de la fuente de contenido ya viene categorizada, por lo tanto, la fase previa de clasificación de contenidos que realizan algunos autores, para este sistema, se puede obviar. La manera con la que se calcula la probabilidad de que al usuario le guste una cierta categoría viene dada por la Ecuación 1. Simplemente se calcula el coeficiente de las veces a las que se accede a una categoría en concreto respecto a las conexiones que realiza el usuario al sistema.

$$Prob(Categoria) = \frac{Accesos\ a\ la\ Categoria}{Accesos\ a\ todas\ las\ categorías}$$

Ecuación 5.1. Probabilidad de una categoría

Una vez obtenida toda la información mediante la web spider, se pasa a la indexación de todo el contenido. Cada noticia extraída pasa por unos ciertos procesos de tratamiento. Primero, se convierte en un objeto que contiene, por separado, el titular, el resumen de la noticia y todo el contenido de la misma. Seguidamente, mediante librerías de Lucene, se indexa, diferenciando los campos de cada bloque de la noticia, y se guarda el resultado en un directorio único para cada análisis completo de toda la fuente de información. Después, se analiza cada bloque para obtener los pesos de sus términos mediante los analizadores de lectura que proporciona Lucene y la fórmula Ecuación 2.

$$Peso(Termino) = FrecTitular_{Termino} * Peso_{Titular} + FrecResumen_{Termino} * Peso_{Resumen} + FrecContenido_{Termino} * Peso_{Contenido}$$

Ecuación 5.2. Obtención del peso de un término en una noticia

Cuando se tiene toda los objetos de las noticias bien construidos, se pueden hacer dos cosas, realimentar el perfil del usuario mediante la navegación por el sistema, o bien, cruzar las noticias con el perfil del usuario utilizando una función de similitud.

Peso titular(Wt)	Peso resumen(Wr)	Peso contenido(Wc)
3	2	1

Tabla 5.1. Pesos para cada bloque de las noticias

TITULAR	**El Banco de *España* ve la economía española como la "más rezagada" de la Zona Euro**
	Euro → Wt*frecuencia(*Euro*) = 3 * 1 = 3 ***España*** → Wt*frecuencia(*España*) = 3 * 1 = 3

Figura 5.1. Análisis del titular

RESUMEN	El director general del Servicio de Estudios del Banco de ***España***, José Luis Malo de Molina, aseguró este viernes que la economía española ha sido la "más rezagada" de la Zona **Euro**, y apuntó que "todavía" existen previsiones de que continúe esta dinámica, ya que se prevén cifras "negativas" en 2010.
	Euro → Wc*frecuencia(*Euro*) = 2 * 1 = 2 ***España*** → Wc*frecuencia(*España*) = 2 * 1 = 2

Figura 5.2. Análisis del bloque de resumen

CONTENIDO	El director general del Servicio de Estudios del Banco de ***España***, José Luis Malo de Molina, aseguró este viernes que la economía española ha sido la "más rezagada" de la Zona **Euro**, y apuntó que "todavía" existen previsiones de que continúe esta dinámica, ya que se prevén cifras "negativas" en 2010. Agregó, durante un coloquio en Las Palmas de Gran Canaria, que mientras en Estados Unidos y Europa en la segunda mitad de 2009 se estaban dando datos positivos, en ***España*** continúan las cifras negativas. "Está rezagada", señaló. A nivel mundial, Malo de Molina afirmó que se ven "síntomas" de que la economía empieza "ha recuperarse", donde sí hay perspectivas de "crecimiento" para 2010. El director general del Servicio de Estudios del Banco de ***España*** que analizó la crisis económica, señaló que en ***España*** ha tenido un carácter dual, ya que le ha afectado a nivel internacional pero también tiene un componente "interno" motivado por los "desequilibrios" que se produjeron en la fase de expansión de la economía española previa a la situación actual. En concreto, indicó que en ***España*** el componente interno que le afectó en la crisis fue, principalmente, el sector inmobiliario, ya que en este ámbito se produjo un "exceso" tanto en los precios de la vivienda como en la inversión o el endeudamiento. Por ello, insistió en la necesidad de que se produzca una "corrección". Todo ello, apuntó, influye en la familia porque ha "debilitado su capacidad de gasto", es decir, del consumo; sin embargo, ha aumentado el ahorro donde, dijo, se están dando tasas "muy altas". En el caso de las empresas, señaló que en estas ha provocado "debilidad", ya que los Beneficios del **Euro** en las mismas han pasado de caer en 2008 un 18 por ciento a alrededor de un 22-23 por ciento el pasado año.Malo de Molina se refirió a la política fiscal como "otro gran instrumento" para reactivar laeconomía. En este sentido, indicó que en ***España*** ha sido una a las que "más se ha recurrido"[...]
	Euro → Wc*frecuencia(*Euro*) = 1 * 2 = 2 ***España*** → Wc*frecuencia(*España*) = 1 * 5 = 5

Figura 5.3. Análisis del bloque de contenido

En las figuras siguientes (5.1, 5.2 y 5.3) se muestra un ejemplo del análisis por cada uno de los bloques de una noticia de economía para un par de términos en concreto, en este caso, Euro y España. Si se presuponen que el peso de los términos en el titular, en el resumen es de 2 y en contenido son los que aparecen en la tabla ,el peso final de Euro corresponderá a 7 y el peso final para el término España será 10.

5.1.1 Funciones de similitud

La clasificación en 'interesante' o 'no interesante' se ha de realizar cruzando los contenidos y el perfil de cada usuario. Como función de similitud inicialmente se ha optado por el Modelo de Espacio Vectorial (MEV). La ecuación del MEV se ha adecuado para esta investigación como se puede apreciar en la ecuación.

Dado un perfil *P* y una noticia N_i, siendo la función *peso(Término)* la encargada de obtener el peso del *Término*, se tiene que la similitud entre *P* y N_i es:

$$SIMIL(P, N_i) = \frac{\sum peso(TPerfil_{j,q}) \cdot peso(TNoticia_{j,i})}{\sqrt{\sum peso^2(TPerfil\ j, q)} \cdot \sqrt{\sum peso_1^2(TNoticia_{j,i})}}$$

Ecuación 5.3. Similitud entre una noticia Ni y el perfil del usuario P

Pero sólo con este resultado para hacer el ranking de todas las noticias a la hora de recomendar las noticias a los usuarios no se ha considerado suficiente. Hay que tener en cuenta la probabilidad del agrado de las categorías. Si sólo se tiene en cuenta el valor de similitud anterior puede darse el caso de que una noticia de una categoría con baja probabilidad tenga un valor de similitud mayor que otra noticia de una categoría de probabilidad alta.

Para resolver esto, la ecuación anterior del modelo de espacio vectorial se modifica como aparece en la ecuación. Manteniendo el planteamiento de la ecuación anterior e insertando la función *Prob(Categoría)*, que es la encargada de obtener la probabilidad de la categoría a la que pertenece la noticia, se tiene que la similitud entre un perfil P y una noticia N_i es como sigue:

$$SIMIL(P, N_i) = Prob(N_i.getCategoria()) \cdot \frac{\sum peso(TPerfil_{j,q}) \cdot peso(TNoticia_{j,i})}{\sqrt{\sum peso^2(TPerfil\ j, q)} \cdot \sqrt{\sum peso_1^2(TNoticia_{j,i})}}$$

Ecuación 5.4. Función de similitud final entre una noticia Ni y el perfil del usuario P

De esta forma, como se verá en el capítulo de evaluación, se obtiene una precisión bastante elevada en los resultados generales.

5.2 Adaptación implícita del sistema a partir de la navegación

Una vez definido cómo se trata y formatea una noticia para su procesamiento y las funciones de similitud que se usan para determinar si una noticia es clasificada como 'interesante', queda describir el aprendizaje del perfil del usuario, cómo va cambiando de forma dinámica a medida que el usuario utiliza y converge con sus gustos.

Se ha definido que el perfil tiene que aprender dos cosas distintas. Por un lado, los contenidos para la clasificación de información, que es lo que se centra esta investigación, pero, como un objetivo secundario, por otro lado debe de aprender a organizar la vista en el dispositivo móvil acorde con la navegación frecuente que tiene el usuario.

5.2.1 Adaptación a partir de los contenidos

Como se ha visto anteriormente, cada noticia pertenece una categoría y finalmente se representa como un vector de elementos clave valor, <*termino, pesoTotal*>. Por otra parte, el perfil del usuario, puede o no tener esa categoría, si la tuviera, tendría asociada una tabla con los términos representativos de esa categoría con sus pesos determinados. El aprendizaje consiste básicamente en agregar la categoría de la noticia a la que se accede si no aún no está en el perfil con su peso correspondiente, y '*fusionar*' el vector de elementos de la noticia con la tabla de los términos de la categoría a la que pertenece la noticia. Pero para ello se van a diferenciar dos etapas, la de inicialización y la de realimentación.

Con el fin de que se entienda mejor, se considerará el perfil de un usuario, denominado *usuarioPrueba*, sobre el que se mostrarán, mediante ejemplos, los contenidos que se explican a continuación. La creación del perfil de *usuarioPrueba* parte con el conjunto de categorías y la tabla con los términos de cada categoría vacíos como se muestra en la figura 5.4.

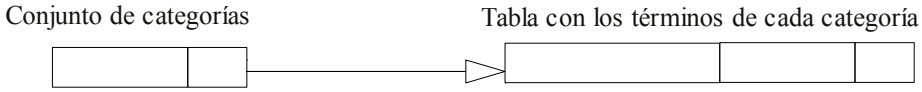

Figura 5.4. Representación del perfil de usuarioPruena

5.2.1.1 Inicialización

Esta etapa se considera el primer acceso del usuario al sistema. Se parte de un perfil vacío, es decir, la tabla de categorías no tiene ninguna entrada por lo tanto no existen relaciones con la tabla de términos, como ya se ha mostrado anteriormente.

Al no existir ningún tipo de información previa, a la hora de adaptar información se han barajado tres posibilidades. Antes de describirlas hay que destacar que el usuario siempre tiene acceso a toda la información de la fuente de contenido en su dispositivo y mediante una navegación completa es capaz de acceder a todas las noticias que se extraen de la fuente.

- La primera consiste en mostrar todas las categorías ordenadas en orden alfabético de las que se dispone para que el usuario acceda a la que esté más interesado. Esta opción tiene el inconveniente que inicialmente el usuario tenga que navegar demasiado para encontrar las noticias que le son de más interés, pero por otro lado, tiene acceso a toda la información que se dispone de la fuente de contenidos desde un principio.

  Por ejemplo, si usuarioPrueba es un forofo del 'Motor' y de los 'Deportes', navegará por estas categorías y accederá a los temas que más le interesan de cada una. Si se supone que en el motor sobre todo le interesan los vehículos de alta gama y en los 'Deportes' su interés principal es el fútbol podría tener un perfil inicial como el de la figura 5.5.

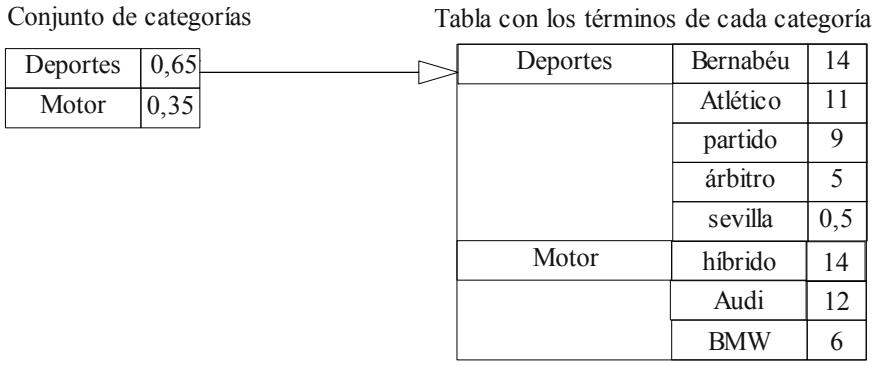

Figura 5.5. Primera posible inicialización del perfil de usuarioPrueba

- La segunda posibilidad es la recomendación de noticias de manera pseudo-aleatoria, es decir, con un cierto azar. Dentro de esta opción, se han considerado otros dos caminos. El más sencillo consiste en que con todo el conjunto de noticias obtenidos en la extracción de la fuente de contenidos, se seleccionen al rededor de unas diez noticias y se muestren al usuario. Este camino tiene el inconveniente de tener muy malos resultados si no se acierta con ninguna a pesar de haber alguna categoría que al usuario le agrade, pero el tema de la noticia no sea de su interés, sin embargo, siempre suele haber un cierto porcentaje de agrado. El otro camino que queda por describir consiste en que, si existen otros usuarios registrados en el sistema, adaptar la información en base de todos los demás usuarios. El problema que puede surgir es que puede haber una gran diferencia respecto los gustos del usuario con el grupo que usuarios, pero para algunos usuarios puede ser positivo de manera que pueden conocer otras áreas y temáticas del mundo real y qué es lo que más se lee.

  En el caso del *usuarioPrueba*, si se toma el caso más sencillo, si existen muchas categorías, con bastante probabilidad inicializará su perfilde forma que no haga una correcta recomendación inicial. Por otro lado, si se toma la segunda opción, no se sabe con seguridad, como está anteriormente comentado, si el grupo de usuarios tienen de alguna forma intereses similares.

- Por último, es mostrar al usuario las últimas diez noticias que se han publicado en la fuente de contenidos. El problema de esto es que si una categoría que no agrada para nada al usuario es la que más se actualiza en la fuente, el usuario lo tomará como una mala recomendación, sin embargo, como en los casos anteriores, esto ayuda a no hacer un foco cerrado del perfil y proponer cosas nuevas e innovar los gustos del usuario.

  Esta opción es la que más suele agradar, ya que las noticias recomendadas de esta manera suelen reflejar los últimos acontecimientos de todas las categorías por lo que, al usuarioPrueba, puede acertar en sus intereses y/o despertar otros nuevos.

En el momento de tomar una elección a la hora de la implementación de un prototipo se llego a la conclusión de que, para ver la repercusión en cada caso, se codificaran todos los casos y que se parametrizara la forma en inicializar los perfiles de los nuevos usuarios.

5.2.1.2 Adaptación a lo largo del tiempo

Teniendo un perfil construido y con una o más conexiones al sistema ya se tiene un cierto modelo de usuario que enfrentado a todo el conjunto de noticias, debe de dar como resultado un convergencia de información de interés. El concepto de 'fusión' se ha comentado anteriormente, que trata de unir el vector de elementos de la noticia con la tabla de los términos de la categoría a la que pertenece la noticia.

Esta unión se lleva a cabo siguiendo el algoritmo de Rocchio que se basa en el supuesto de que la mayoría de los usuarios tienen una percepción general de que los documentos se denotan como relevantes o no relevantes.

$$\vec{Q}_m=(a\cdot\vec{Q}_o)+(b\cdot\frac{1}{|Dr|}\cdot\sum\vec{Dr}_j)-(c\cdot\frac{1}{|Dnr|}\cdot\sum\vec{Dnr}_k)$$

Ecuación 5.5. Fórmula original de Rocchio

La ecuación anterior es la general de Rocchio (ecuación 5.5) , pero para el este estudio se ha modificado con la novedad de considerar las consultas anteriores (Qm y Qo) como el perfil del usuario (Pm y Po), los pesos a y b con valor 1 y el peso de c con valor 0:

$$\vec{P}_m=(1\cdot\vec{P}_o)+(1\cdot\frac{1}{|Dr|}\cdot\sum\vec{pesos}\cdot\vec{Dr}_j)-(0\cdot\frac{1}{|Dnr|}\cdot\sum\vec{Dnr}_k)$$

Ecuación 5.6. Ecuación de Rocchio modificada

Si se simplifica y se tiene en cuenta que se realimenta al perfil por cada noticia a la que el usuario accede, el término |Dr| = 1. Finalmente se obtiene:

$$\vec{P}_m=(\vec{P}_o)+(\sum\vec{Dr}_j)$$

Ecuación 5.7. Ecuación de Rocchio modificada y simplificada

La 'fusion' entre una noticia y el perfil de un usuario no sólo consiste en añadir términos nuevos y sumar los pesos, también necesita perder algún tipo de información para evitar el ruido y hacer frente a los cambios de interés. Esto se ha querido llevar a cabo aplicando un 'factor de olvido' *f*.

Este factor se encarga de que los términos que dejen de interesar al usuario, o bien, realmente no sean de utilidad desaparezcan del perfil del usuario. El valor de este factor se ha considerado que se encuentre entre los valores mínimo y máximo de los pesos que se proporciona a cada bloque de las noticias.

Una regla importante que se ha establecido en el sistema consiste en que todo término que tenga un valor cero o menos que cero es eliminado del perfil del usuario. Esta medida se toma, por un lado, para que los números negativos no repercutan de manera contraproducente en el cálculo de similitud, y, por otro, para que el perfil del usuario no incremente de manera excesiva su tamaño y su tratamiento sea lo más ágil y viable posible.

A continuación se muestra un ejemplo de cómo se realimenta el perfil de usuarioPrueba, que ha accedido a una noticia que le interesa que pertenece a la categoría 'Deportes'. Se supone que los términos que forman parte de las noticias y el

perfil son las palabras del texto. Además, se toman los pesos del titular, el resumen y el contenido como 4, 3 y 1, respectivamente. Un fragmento de la representación de esta noticia de deportes con los términos más significativos es la que se muestra en la tabla 5.2.

Término	**Valor**
atlético	16
europa	13
forlán	12
madrid	11
uefa	4
rojiblanco	3
league	4
afición	3

Tabla 5.2. Términos más significativos de uno noticia de interés.

Tomando el vector de términos de la categoría 'Deportes' del perfil de *usuarioPrueba* que se ha obtenido en la primera inicialización, se realiza la 'fusión' con el contenido de la noticia. El resultado de esta fusión se muestra en la tabla 6.

Si se analiza con más detalle se puede ver que al término 'atlético' se le ha sumado el valor con el que aparece en la noticia.

Los términos que no aparecen en esta noticia se les resta *f*, el factor de olvido que se establece en forma de parámetro, como se ha comentado antes, depende de la capacidad de adaptación que se considere frente a cambios de intereses por parte del usuario, que en este caso toma el valor de 1, es decir, el mínimo peso de los tres bloques de una noticia. También se puede ver que aparece el caso en el que un término ha de desaparecer del perfil, como es el término 'sevilla', que aparece tachado y tiene un peso en el perfil de 0,5 y al pasarlo el factor de olvido es menor que cero, con lo que automáticamente se ha de quitar.

Los últimos términos de la tabla son los pertenecientes a la noticia que es de interés para el usuario, estos términos junto a su valor se añaden sin más al perfil, para futuras comparaciones de similitud.

Término	Valor
Bernabéu	14 – *f*
atlético	11 + 16
partido	9 – *f*
árbitro	5 – *f*
~~sevilla~~	~~0,5 - f~~
europa	13
forlán	12
madrid	11
uefa	4
rojiblanco	3
league	4
afición	3

Tabla 5.3. Vector final de la categoría 'Deportes' de usuarioPrueba

5.2.2 Adaptación a partir de la navegación

Desde la introducción de este trabajo se ha hablado de una construcción del perfil completamente implícita mediante la navegación en el dispositivo móvil. A continuación se describe en qué consiste esta navegación y qué aporta cada acción del usuario.

Navegación por las categorías

Las categorías se van ordenando a medida que el usuario va accediendo a cada una de ellas. Son la raíz de todo la información que se extrae de la fuente de contenidos, por lo tanto se deben de facilitar al usuario todas las existentes, aunque tenga una probabilidad de acceso cero, que, en ese caso, se encontrarán en la última posición de la lista como se muestra en la figura 5.6.

El número de categorías es variable, siendo definido previamente, por un lado, en el servidor por el administrador del sistema, y por otro, las que pueda proveer la fuente de contenidos. Cuando un usuario accede a las noticias de una categoría en concreto de forma automática alimenta al perfil del usuario modificando la probabilidad de acceso.

Figura 5.6. Categorías personalizadas al usuario

Navegación por las noticias

Se tienen dos conjuntos de noticias, las que están recomendadas por el sistemas y las que dependen de una categoría en concreto. En ambos el sistema es capaz de obtener información para enriquecer el perfil del usuario

Las primeras dependen de forma directa del perfil del usuario, es decir, son las que el sistema propone al usuario nada más conectarse con su dispositivo móvil (Figura). Estas noticias están ordenadas según la similitud que se tiene con el perfil, de forma que la qué mas se espera que interese al usuario se encuentre en primer lugar, pudiendo incluso mostrara el contenido pleno de la noticia si supera un cierto mínimo de similitud. La cantidad de noticias recomendadas es variable, siendo el intervalo más adecuado entre 10 y 20 noticias para un dispositivo móvil como se ha demostrado en el capítulo de evaluación.

Figura 5.7. Conjunto de noticias recomendadas

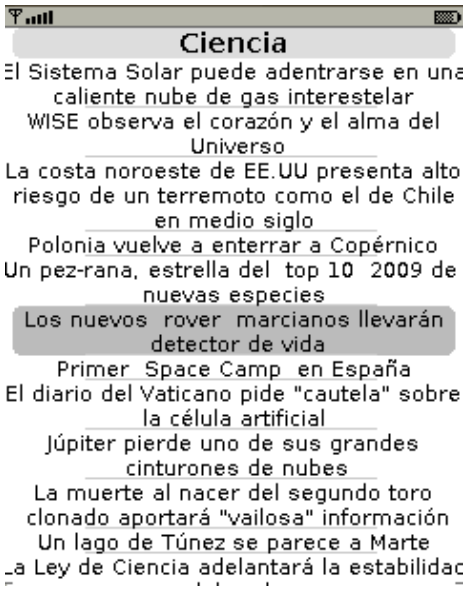

Figura 5.8. Conjunto de noticias dependientes de una categoría

Las noticias dependientes de una categoría son accedidas mediante el acceso a la misma (Figura). La cantidad de noticias depende de las que proporciona la fuente en ese día en una primera consulta, sin embargo, existe la posibilidad de acceder a cualquier

noticia de un algún día anterior. También tienen orden diferente para cada usuario intentando ajustar el nivel de similitud con su perfil, probablemente sea mínimo o incluso cero, en ese caso se muestran por orden cronológico.

Acceso pleno a una noticia

El acceso pleno a una noticia tiene dos acciones. Primeramente, es mostrar todo el contenido de la misma, es decir, los tres bloques, titular, resumen y contenido. Segundo, alimentar al perfil del usuario con la noticia accedida con la metodología explicada en las secciones anteriores.

Puede aparecer el caso en el que un usuario nunca acceda a al contenido pleno de una noticia, es decir, que sólo le interese consultar los titulares sin llegar a este punto. Puede ser un problema ya que si nunca accede al contenido de una noticia, ¿cómo se puede conseguir una personalización de la nada?. Para ello se han barajado algunas soluciones:

- Información temporal o auxiliar en el perfil por donde pasa el foco el usuario, se enriquece el perfil del usuario con los titulares por donde pasa el foco en las listas de los conjuntos.
- Establecer un tiempo mínimo en donde el foco en la lista del conjunto de noticias dependientes de una categoría o de las que se recomiendan se encuentra sin moverse, de forma que, pasado este tiempo, se interpreta que se está leyendo el titular y puede agregarse para enriquecer el perfil, descartando el resto.


Ciencia

Un lago de Túnez se parece a Marte

En los trabajos de campo –que se llevarán a cabo entre este lunes, 24 de mayo, y el próximo día 31 de mayo– participarán expertos en geología, geobiología, sedimentología y geología planetaria que forman parte del programa Europlanet de la Unión Europea.

En los trabajos de campo --que se llevarán a cabo entre este lunes, 24 de mayo, y el próximo día 31 de mayo-- participarán expertos en geología, geobiología, sedimentología y geología planetaria que forman parte del programa Europlanet de la Unión Europea. La campaña está dirigida por el investigador Felipe Gómez, del Centro de Astrobiología, centro mixto del CSIC y del Instituto Nacional de Tecnología Aeroespacial. Según el experto, "la mayor parte del año el lago presenta

Figura 5.9. Acceso pleno a una noticia.

6 Caso de uso de la aplicación en el dispositivo

Dentro del uso de la aplicación en el dispositivo, ésta consta de cuatro partes o secciones generales. La sección de categorías, la sección de noticias de una categoría concreta, la sección de noticias recomendadas y el acceso pleno a una noticia. En todo momento, al usuario al usuario se le permite llegar a todas estas partes mediante accesos directos.

6.1 Sección de categorías

Esta sección aparece en dos casos y proporciona al usuario el inicio para explorar la totalidad de las noticias de la fuente de contenido. Las categorías están dispuestas en una maya que se ordena dependiendo de las preferencias del usuario, colocándose para el final las opciones de configuración y los créditos.

Los dos casos en los que se accede a esta sección pueden ser la siguientes:

- Si el usuario es la primera vez que accede al sistema, tiene un perfil vacío, por lo tanto no es posible recomendarle nada, de forma que se le redirecciona a esta sección.
- Si el usuario se encuentra en cualquier sección y desea empezar desde el principio a explorar otra sección.

Figura 6.1. Captura de la sección de categorías

6.2 Sección de noticias recomendadas

En esta parte aparecen las noticias que el sistema considera que son de interés para el usuario. Pueden aparecer noticias de cualquier categoría colocadas en función de la similitud que ha estimado el sistema.

Esta sección se le ofrece al usuario si ya antes había accedido a algún tipo de contenido, es decir, su perfil tiene alguna información, nada más acceder a la aplicación de su dispositivo. Siempre puede volver a ella pulsando el botón de recomendadas que aparecerá en otras secciones.

Esta sección va cambiando a medida que el usuario accede a las diferentes secciones de la aplicación y a los contenidos asociados.

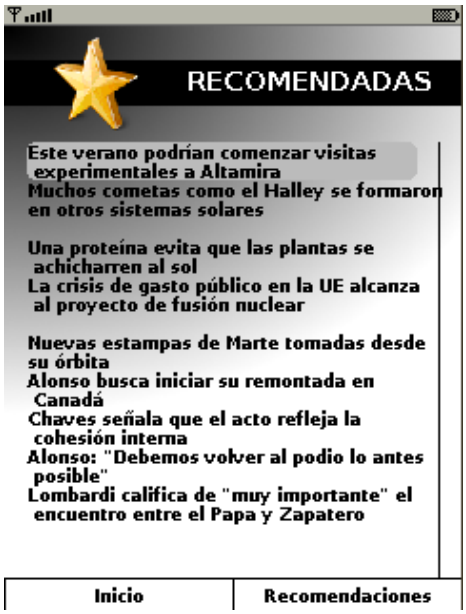

Figura 6.2. Sección de noticias recomendadas.

6.3 Sección de noticias de una categoría

A esta sección se puede acceder desde la sección de categorías. Cuando el usuario pulsa sobre un icono en concreto, el dispositivo muestra todas las noticias de la categoría. Cada instancia de esta sección tiene propiedades propias, como el fondo de pantalla y el icono de la categoría que representa.

El sistema intenta realizar la ordenación de estas noticias acorde al perfil del usuario, las noticias son iguales para todos los usuarios, pero varía el orden en que se presentan.

Figura 6.3 Distintas instancias de la sección de las noticias de una categoría en concreto

6.4 Sección de una noticia completa

Acceder a esta sección es lo que realmente enriquece el perfil del usuario. Se puede llegar hasta ella desde la sección de noticias recomendadas o bien desde la de noticias de una categoría. Únicamente consiste en mostrar la noticia completamente, es decir, los tres bloques de titular, resumen y contenido, e información multimedia si se dispusiera de ella.

Desde aquí se puede volver a la sección de categorías o bien a las recomendadas para comprobar qué hay de nuevo en la fuente de contenidos que pueda interesar al usuario.

Figura 6.4. Sección dónde aparece una noticia completa

IV. EVALUACIÓN

1 Introducción

Cuando se ha querido hacer una evaluación exhaustiva del sistema y de los elementos utilizados, se ha planteado la cuestión del comportamiento con usuarios ideales y usuarios reales. Para llevar a cabo la evaluación con usuarios ideales no se requiere demasiado trabajo, carecen de ruido y se pueden obtener resultados bastante concluyentes en situaciones idílicas. Por otro lado, comprobar el comportamiento con usuarios reales proporciona un conjunto de datos más útiles y tienen más peso a la hora de concluir los resultados.

Partiendo la evaluación con perfiles ideales, se van a diferenciar dos tipos, un perfil tipo equilibrado (perfil 2) y un perfil tipo focalizado (perfil 1). Por un lado, el perfil 1, que tendrá una predilección dominante por una de las categorías, habiendo otras dentro del mismo, pero con una probabilidad muy baja y acompañando cada categoría, un conjunto términos en el que cada uno tendrá su peso. Por otro lado, el perfil 2, tendrá una predilección por varias categorías de una forma más equilibrada y ,como en el tipo de perfil anterior, una conjunto de términos con sus pesos respectivos.

Referente a los usuarios reales, se ha tenido la colaboración de diez personas de diferentes ámbitos. Personas dedicadas a la investigación en informática o en medios de comunicación, funcionarios, amas de casa o entrenadores personales. Como se verá cada usuario tendrá un perfil muy diferente y en algunos casos no se ajustarán con ninguno de los casos ideales comentados anteriormente.

La evaluación del sistema se realizará a lo largo del tiempo, es decir, durante varios días, aunque con algunas diferencias entre los casos ideales y los reales. Para los perfiles ideales, se hará solamente en dos días. El primer día se toma como inicial, donde cada usuario navega y ve las noticias que le parecen mas adecuadas. El segundo día es donde se compara cómo de buena ha sido la representación de los gustos del usuario con respecto a sus preferencias en realidad.

Por otro lado, para el conjunto de usuarios reales se toma un período de siete días, de esta manera se puede observar la evolución de la representación de los perfiles de cada usuario y la progresión de clasificación de los contenidos.

Para ambos casos, la evaluación de la personalización de contenidos que hace el sistema se realizará desde cuatro enfoques distintos. Primero, tomando el peso de todos los términos del mismo valor, es decir, sin tener en cuenta el bloque de la noticia en el que aparecen. Segundo, con pesos variables, es decir, dando el peso correspondiente teniendo en cuenta la aparición de un cierto término en un bloque determinado. Estos dos enfoques con todo el contenido que proporciona una noticia. Tercero, teniendo en cuenta únicamente el titular y el resumen de la noticia con los pesos variables. Y por último, se utiliza únicamente la información proporcionada por el contenido, sin el titular ni el resumen, de una noticia con pesos variables.

En cada enfoque se ha ramificado las pruebas en dos vertientes más, por un lado usar las palabras sin más, y por otro, usando un analizador de raíces. Para ambos casos se excluyen las palabras vacías como preposiciones, adverbios, etc.

En resumen, este capítulo consiste en un plantear la evaluación describiendo los parámetros que se tienen en cuenta y combinarlos para obtener diferentes resultados.

2 Planteamiento

2.1 Métricas

Para medir los resultados que se obtendrán en este capítulo se va hacer uso de tres métricas distintas bastante usadas en el campo de la recuperación y recomendación de información. A continuación se describen cada una de estas métricas:

- MAP, Mean Average Precision, Precisión y recall son métricas de un único valor sobre que se basa de todo un conjunto de documentos devueltos por el sistema. Para los sistemas que devuelven una secuencia ordenada de los documentos, es conveniente tener en cuenta también el orden en que se presentan estos documentos devueltos. Este tipo de media hace hincapié en los documentos en la parte superior del ranking. Es el promedio de la precisión calculada en el punto de cada uno de los documentos pertinentes en la secuencia de clasificación. La ecuación que la define es la siguiente:

$$MAP = \frac{\sum (P(r) \cdot rel(r))}{número\,de\,documentos\,relevantes}$$

Ecuación 2.1. Fórmula para el cálculo de la MAP

La función rel(r) es sencillamente una función binaria que indica si la noticia de la posición r del ranking es relevante o no. Por otro lado, P(r) es la precisión dada por la fórmula siguiente:

$$P(r) = \frac{|documentos\,relevantes\,hasta\,una\;posición\,r\,del\,ranking|}{r}$$

Ecuación 2.2. Fórmula para el cálculo de la precisión en un punto r del ranking

- P10, Con este valor se pretende medir la precisión sobre los 10 primeros documentos recomendados, ya que en un dispositivo móvil es lo que se recomendará a los usuarios inicialmente. Se puede reutilizar la fórmula de la precisión anterior pero dando a *r* el valor 10.

$$P(10) = \frac{|documentos\,relevantes\,hasta\,la\;posición\,10\,del\,ranking|}{10}$$

Ecuación 2.3. Fórmula para el cálculo de la precisión en los 10 primeros noticias del ranking

- P15-20, Con este valor se pretende medir la precisión entre 15 y 20 primeros documentos recomendados, de esta manera se intenta estudiar si, a pesar de recomendar más noticias, como se comporta la precisión. Como en la métrica anterior se puede reutilizar P(r), aunque en este caso se hará una media entre ambas para simplificar los resultados.

$$P(15{,}20) = \frac{P(15) + P(20)}{2}$$

Ecuación 2.4. Precisión en las 15 y 20 noticias primeras

2.2 Contenidos

El conjunto de noticias con las que se cruzan los perfiles son para las cuatro enfoques el mismo, por lo tanto, se obtienen unos datos consistentes. Para cada perfil se usan los conjuntos de todas las noticias de las categorías que aparecen en ellos. El resto de categorías no son necesarias porque se supone que no interesan al usuario.

Para cada día de evaluación se obtiene el conjunto de noticias correspondiente, no todos los días existen el mismo número de noticias. Tanto en la tabla 7 como en la figura se puede observar la relación de obtención noticias con respecto al tiempo.

	Día 1	**Día 2**	**Día 3**	**Día 4**	**Día 5**	**Día 6**	**Día 7**
Número	103	104	87	82	88	106	105

Tabla 2.1. Número de noticias extraídas al día

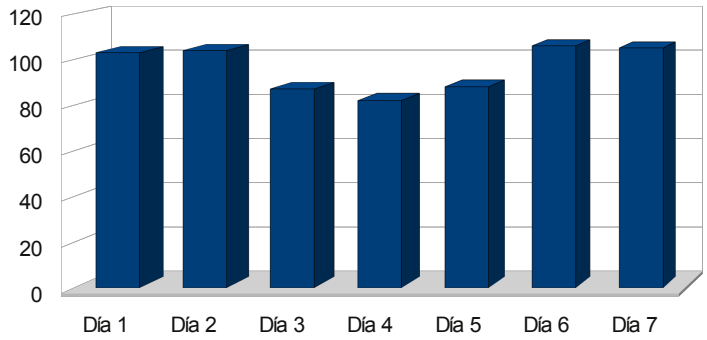

Figura 2.1. Noticias obtenidas cada día de la semana

Dentro de cada día se tiene un total de nueve categorías, que son las que aparecen en la tabla.

Categorías
Ciencia
Cultura
Deportes
Economía
Motor
Noticias Internacionales
Noticias Latinoamericanas
Noticias Nacionales
Televisión

Tabla 2.2. Relación de las categorías contempladas

La frecuencia de noticias tampoco es siempre constante para cada categoría. Tampoco existe un patrón sobre en qué día aparecen más noticias de una cierta categoría dentro de una sola semana. En la siguiente figura se muestra un gráfico más detallado de la proporción del número de noticias obtenidas para cada uno de los siete días.

Se ha de aclarar que, a pesar de que una cierta categoría tenga una mayor

probabilidad de gustar al usuario, el número de noticias recomendadas de esa categorías no tiene que ser necesariamente proporcional a la misma. Si existen noticias de otras categorías de menor probabilidad, pero en cambio, tienen un peso mayor que una noticia de la categoría predominante, se clasificará como interesante para el usuario, de esta manera se pretende conseguir un equilibrio e intentar que el usuario tenga la sensación de que el sistema lo conoce lo suficiente.

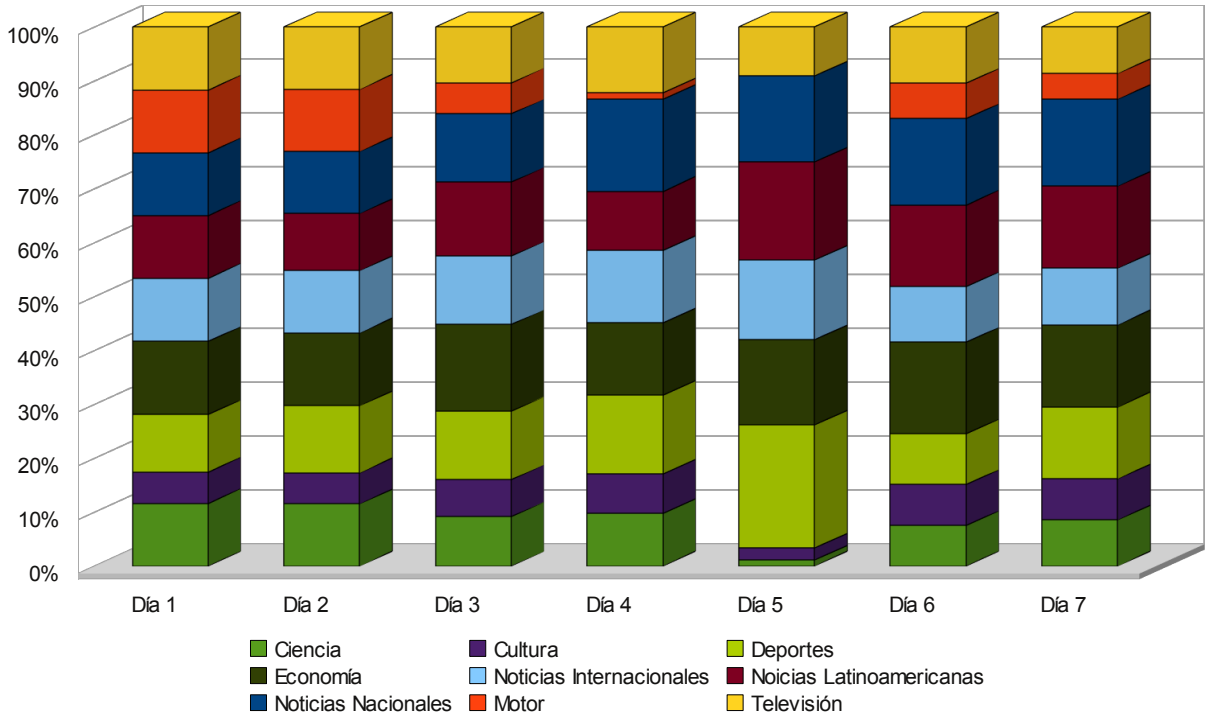

Figura 2.3. Proporción al día del número de noticias por categorías

2.3 Perfiles ideales

Los perfiles ideales se han construido a mano para comprobar su fiabilidad y unas conclusiones iniciales.

Perfil 1			Perfil 2		
Categoría	**Prob**	**Palabras clave**	**Categoría**	**Prob**	**Palabras clave**
“deportes”	0,8	Fútbol (0,99)	“cultura”	0,4	Cine (0,99)
		Madrid (0,67)			Arte (0,75)
		Partido(0,53)			Historia (0,5)
		Tenis (0,46)			Musica (0,4)
“motor”	0,1	Ferrari (0,99)	“economía”	0,3	Banco (0,99)
		Mclaren (0,73)			Euribor (0,8)
		Fernando (0,5)			Negocio (0,66)
“nacional”	0,1	Gobierno(0,99)	“salud”	0,3	Estudio (0,99)
		Zapatero(0,77)			Cancer (0,7)
		Rajoy(0,69)			Vacuna (0,65)

Tabla 2.3. Instancia de dos perfiles personalizados

Para facilitar esta parte de la evaluación, una instancia sencilla de estos dos perfiles se muestran en la tabla 2.3.Como se ha comentado con anterioridad, la construcción de estos perfiles se ha llevado acabo en dos días, es decir, con los dos primeros días que se muestran en el planteamiento de los contenidos. El primero donde se va creando el perfil y el segundo, en le que se recomiendan las noticias y se comparan con las que realmente le interesarían a estos usuarios ideales.

Por lo tanto, para esta evaluación, en el perfil 1, se tomarán el conjunto de noticias de las categorías 'deportes', 'motor' y 'nacional' y, en el perfil 2, se tomarán el conjunto de noticias 'economía', 'cultura' y 'salud'.

2.4 Perfiles de usuarios reales

Para obtener una evaluación más acorde con la realidad, durante una semana se ha llevado a cabo un conjunto de encuestas a diez personas sobre las noticias que le interesan de cada día. El formato y un ejemplo de estas estas encuestas se encuentra en el ANEXO I.

De la misma manera que cada día hay un número diferente de noticias, a los usuarios les puede interesar un número diferente de noticias. La tabla 2.4 que aparece a continuación contiene el número de noticias de interés de cada usuario al día y la última columna muestra el porcentaje total de las noticias que le han interesado al usuario en toda la semana.

	Día 1	**Día 2**	**Día 3**	**Día 4**	**Día 5**	**Día 6**	**Día 7**	**% Interés**
Usuario 1	20	20	15	17	10	23	17	18,07%
Usuario 2	53	39	32	40	32	54	52	44,74%
Usuario 3	81	77	42	43	47	57	59	60,14%
Usuario 4	52	41	34	37	37	35	25	38,66%
Usuario 5	54	57	51	29	27	36	49	44,89%
Usuario 6	27	19	10	16	8	8	21	16,14%
Usuario 7	59	60	63	58	76	67	75	67,85%
Usuario 8	15	16	7	6	6	11	11	10,67%
Usuario 9	39	50	33	18	12	23	18	28,59%
Usuario 10	25	22	11	9	11	14	21	16,74%

Tabla 2.4. Relación de noticias de interés para cada usuario por día de forma absoluta y en proporción al total de noticias

Con lo datos de estos usuarios se pueden sacar patrones con los que agruparlos en conjuntos para su estudio. Esta agrupación depende de la proporción de las noticias que están interesados al día con las que existen en ese día. Los grupos que se pueden identificar son los siguientes:

- Grupo de usuarios con interés informativo bajo, en este grupo se encuentran los usuarios que de todas las noticias que se les propone o están disponibles en un día, están interesados en menos del 30 %. Pueden hacer un uso intensivo del sistema, es decir, hacer muchas conexiones, o bien, de conectarse y consultar las noticias de manera eventual.

- Grupo de usuarios con interés informativo alto, este grupo de usuarios suelen ser los más activos. Aquí se encuentran los usuarios que están interesados en más del 60 % de las noticias existentes en un día.
- Grupo de usuarios con interés informativo medio, este último grupo es variable y sus intereses a lo largo de la semana pueden tener picos y depresiones, o bien, sus las conexiones al sistema son cortas pero frecuentes. En unas ocasiones pueden alcanzar el 80 % de noticias que les interesan y en otras no llegar al 20 %. Suele ser el más frecuente.

En la tabla 2.5 se muestran los usuarios que se identifican en cada grupo:

Interés bajo	**Interés medio**	**Interés alto**
Usuario 1	Usuario 2	Usuario 3
Usuario 6	Usuario 4	Usuario 7
Usuario 8	Usuario 5	
Usuario 10	Usuario 9	

Tabla 2.5. Identificación de los usuarios según su grupo

En esta evaluación se harán los experimentos para estos grupos tomando el mejor enfoque obtenido de la comparación general, en lugar de para cada individuo con el fin de simplificar los resultados. Aún así, en el ANEXO II se encuentran los datos de cada usuario individualmente y su progreso durante toda la semana para futuras consultas o profundizar en estos resultados.

3 Estudio con perfiles de usuario ideales

3.1 Enfoque de los términos con el mismo peso

En este enfoque se han tomado los pesos de los tres bloques de la noticia con valor a 1.Las noticias que se proporcionan al usuario con perfil 1 son 10, de las cuales le parecen interesantes 4 por lo tanto, la precisión para este enfoque es de 0,4, o bien, del 40%. Por otro lado, para el cálculo del recall, tomando los documentos recuperados relevantes y el total de noticias relevantes para el usuario, que son 14, se obtiene un 0,285, es decir, un 28,5%. En cambio, el usuario con perfil 2, toma como relevantes 4 noticias de las propuestas, obteniendo una precisión de 0,4 (40%). El recall, con 22 documentos relevantes en total dentro del conjunto de las noticias es de 0,182 (18,2%).

Categ. Noticia	Valor	Clase	Categ. Noticia	Valor	Clase
Deportes	0,7	Interesa	Cultura	0,8	Interesa
Deportes	0,42	Interesa	Economía	0,6	No interesa
Deportes	0,4	Interesa	Cultura	0,55	Interesa
Deportes	0,4	No interesa	Economía	0,48	No interesa
Nacional	0,3	No Interesa	Economía	0,47	No interesa
Nacional	0,24	Interesa	Economía	0,44	No interesa
Deportes	0,2	No Interesa	Economía	0,4	No interesa
Nacional	0,2	No interesa	Economía	0,3	Interesa
Nacional	0,1	No interesa	Ciencia	0,1	Interesa
Deportes	0,1	No interesa	Ciencia	0,1	No interesa

	Perfil 1	Perfil 2
Not. Recuperadas	10	10
Not. Recuperadas Relevantes	4	4
Not. Relevantes	14	22
Precisión	0,4	0,4
Recall	0,29	0,18

Tabla 3.1. Resultados tomando el enfoque de pesos unitarios

3.2 Enfoque de los términos con peso variable

Los pesos en este enfoque se han tomado con valor 3 *(Wt = 3)* a los términos del titular, con valor 2 *(Wr = 2)*, a los términos del resumen y con valor 1 *(Wc = 1)* los términos del contenido. Las noticias que se proporcionan al usuario con perfil 1 son 10 de las cuales le parecen relevantes 6, por lo tanto, la precisión para este enfoque es de 0,6 (60%). Por otro lado, para el cálculo del recall, tomando los documentos recuperados relevantes y el total de noticias relevantes, que son 14 para el usuario se obtiene un 0,42 (42 %). En cambio, el usuario con perfil 2, toma como relevantes 7 noticias de las propuestas, obteniendo una precisión de 0,7 (70%). El recall, con 22 documentos relevantes en total dentro del conjunto de las noticias es de 0,32 (32%).

Perfil 1			Perfil 2		
Categ. Noticia	**Valor**	**Clase**	**Categ. Noticia**	**Valor**	**Clase**
Deportes	0,8	Interesa	Economía	0,56	Interesa
Nacional	0,7	Interesa	Economía	0,55	Interesa
Motor	0,6	Interesa	Cultura	0,55	Interesa
Deportes	0,55	Interesa	Economía	0,4	No interesa
Deportes	0,5	Interesa	Ciencia	0,3	Interesa
Nacional	0,54	No Interesa	Cultura	0,1	Interesa
Deportes	0,4	No interesa	Cultura	0,1	Interesa
Nacional	0,34	Interesa	Economía	0,1	Interesa
Deportes	0,35	No interesa	Cultura	0,1	No interesa
Nacional	0,33	No interesa	Economía	0,1	No interesa

	Perfil 1	Perfil 2
Not. Recuperadas	10	10
Not. Recuperadas Relevantes	6	7
Not. Relevantes	14	22
Precisión	0,6	0,7
Recall	0,42	0,32

Tabla 3.2. Resultados tomando el enfoque de pesos variables

3.3 Enfoque tomando el titular y el resumen con pesos variables

Ahora, se mantienen los pesos del enfoque anterior, menos el del contenido, que toma valor cero, de esta manera se excluye de la similitud. Las noticias que se proporcionan al usuario con perfil 1 son 5 de las cuales le parecen relevantes 4, por lo

tanto, la precisión para este enfoque es de 0,8 (80%). Por otro lado, para el cálculo del recall, tomando los documentos recuperados relevantes y el total de noticias relevantes para el usuario se obtiene un 0,285 (28,5%). En cambio, el usuario con perfil 2, toma como relevantes 3 noticias de las 4 propuestas, obteniendo una precisión de 0,75 (75%). El recall, con 22 documentos relevantes en total dentro del conjunto de las noticias. es de 0,136 (13,6%).

Perfil 1			Perfil 2		
Categ. Noticia	**Valor**	**Clase**	**Categ. Noticia**	**Valor**	**Clase**
Nacional	0,7	Interesa	Cultura	0,5	Interesa
Deportes	0,6	Interesa	Economía	0,4	Interesa
Deportes	0,4	Interesa	Ciencia	0,33	Interesa
Deportes	0,3	Interesa	Economía	0,1	No interesa
Nacional	0,15	No interesa			

	Perfil 1	**Perfil 2**
Not. Recuperadas	5	4
Not. Recuperadas Relevantes	4	3
Not. Relevantes	14	22
Precisión	0,8	0,75
Recall	0,28	0,13

Tabla 3.3 Resultados tomando el enfoque del titular y resumen con pesos variables

3.4 Enfoque tomando el contenido con pesos variables

Finalmente, en este enfoque se han tomado los pesos del titular y el resumen de la noticia a cero, así quedan fuera de la similitud, dando al peso de los términos del contenido valor 1. Las noticias que se proporcionan al usuario con perfil 1 en este último enfoque son 9 de las cuales le parecen relevantes 5, por lo tanto, la precisión para este enfoque es de 0,55 (55%). Por otro lado, para el cálculo del recall, tomando los documentos recuperados relevantes y el total de 14 noticias relevantes para el usuario se obtiene un 0,357 (35,8%). En cambio, el usuario con perfil 2, toma como relevantes 4 noticias de las 7 propuestas, obteniendo una precisión de 0,571 (57,1%). El recall, con documentos relevantes en total dentro del conjunto de las noticias es de 0,18 (18%).

Perfil 1			Perfil 2		
Categ. Noticia	**Valor**	**Clase**	**Categ. Noticia**	**Valor**	**Clase**
Deportes	0,7	Interesa	Cultura	0,55	Interesa
Deportes	0,7	Interesa	Economía	0,5	No interesa
Deportes	0,66	No interesa	Economía	0,48	Interesa
Nacional	0,6	Interesa	Cultura	0,4	No interesa
Deportes	0,4	Interesa	Economía	0,1	Interesa
Nacional	0,32	No interesa	Ciencia	0,1	Interesa
Nacional	0,21	No interesa	Cultura	0,1	No interesa
Deportes	0,1	Interesa			
Deportes	0,1	No interesa			

	Perfil 1	**Perfil 2**
Not. Recuperadas	9	7
Not. Recuperadas Relevantes	5	4
Not. Relevantes	14	22
Precisión	0,55	0,57
Recall	0,35	0,18

Tabla 3.4. Resultados tomando el enfoque del contenido con pesos variables

4 Estudio con perfiles de usuarios reales

4.1 Resultados generales

En la tabla 16 se presentan los resultados generales obtenidos para toda la semana en cada uno de los enfoques propuestos en la introducción. Estos resultados generales se han sacado como la media de cada usuario de forma individual. La última fila representa la media de todo el período de experimentación. En la tabla 4.1 se muestra los resultados generales para la métrica MAP y P10.

	Pesos Unitarios (PU)		**Pesos Variables (PV)**		**Titular + Resumen (T+R)**		**Contenido (C)**	
	MAP	**P10**	**MAP**	**P10**	**MAP**	**P10**	**MAP**	**P10**
Día 1	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Día 2	0,6760	0,6900	0,6861	0,6900	0,6949	0,7200	0,6662	0,6900
Día 3	0,6063	0,6400	0,6169	0,7100	0,6286	0,7600	0,6631	0,7000
Día 4	0,5695	0,6500	0,5798	0,7000	0,5809	0,7300	0,5622	0,6300
Día 5	0,5456	0,6400	0,5529	0,6600	0,5491	0,7100	0,5411	0,6400
Día 6	0,5377	0,7400	0,5451	0,7500	0,5339	0,7900	0,5335	0,7300
Día 7	0,5334	0,7700	0,5396	0,7700	0,5300	0,7900	0,5293	0,7600
Media	**0,5780**	**0,6884**	**0,5867**	**0,7133**	**0,5862**	**0,7500**	**0,5825**	**0,6916**

Tabla 4.1. Resultados con las métricas MAP y P10 para los diferentes enfoques

Si tomamos las medias de los MAPs de los cuatro enfoques, como se muestra en la tabla 4.2, se deduce que en un 1,5 %, el enfoque de tomar toda la noticia con pesos variables es mejor que tomar pesos unitarios. Además de mejorar en un 0,7% y un 0,08% a los enfoques que toman el titular y el resumen y el que toma únicamente el contenido.

Comparación	**Porcentaje de mejora**
PV > PU	1,50%
PV > T+R	0,08%
PV > C	0,70%

Tabla 4.2. Comparación usando la métrica MAP

En el caso de tomar los valores de la métrica P10, las mejoras son más significativas. Comparando el T+R con el PU y el C se obtiene una mejora en ambos sobre el 8%. Si se compara T+R con PV, la mejora supera el 5%. Estas comparaciones se ven de forma más clara en la tabla 4.3.

Comparación	Porcentaje de mejora
T+R > PU	8,94%
T+R > PV	5,14%
T+R > C	8,44%

Tabla 4.3. Comparando usando la métrica P10

Después de la comparación de las medias finales de las métricas como aparecen en las tablas anteriores, se deduce que el mejor enfoque es el de tomar el titular y el resumen de la noticia de forma proporcionalmente significativa. Para mostrar el progreso de adaptación de los enfoques para las métricas de MAP y P10, además, con el fin de estudiar la cobertura de la precisión se utiliza la métrica de P15-20. Para cada enfoque se ha realizado una gráfica de su progresión.

En la figura 4.1, se muestra la progresión del enfoque que tiene mejores resultados respecto a las medidas de precisión, que es el de tomar el titular y el resumen. Como se puede ver, ambas medidas de precisión se encuentran en constante crecimiento salvo una pequeña disminución a mediados del período de evaluación. La medida del MAP empeora a lo largo de la semana, pero tiende a ser estable al final de la misma.

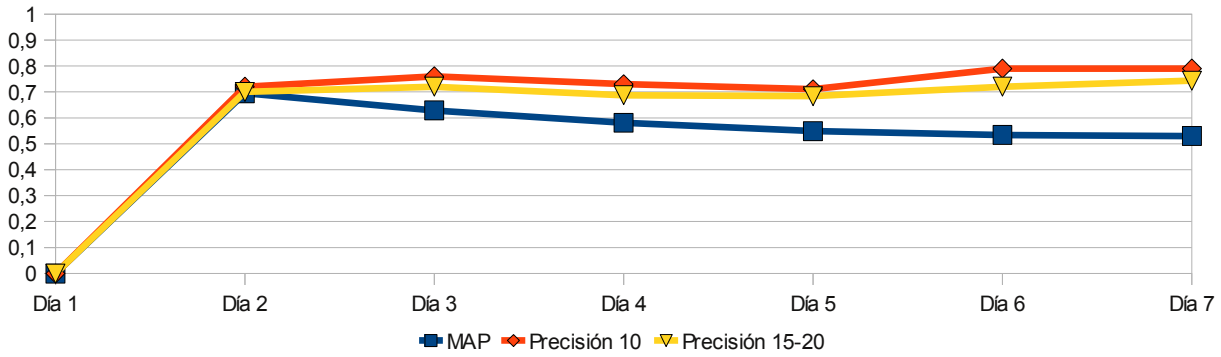

Figura 4.1. Progreso general del sistema durante una semana con el enfoque de tomar el titular y el resumen

Con el segundo mejor enfoque, el de tomar toda la noticia con pesos variables, en la figura 4.2, se obtienen unas características más irregulares, sobre todo con las medidas de precisión, que existen picos y depresiones, pero al final de la semana, estas medidas van en aumento. Como en el caso anterior, la métrica MAP disminuye, hasta mantenerse estable los últimos días.

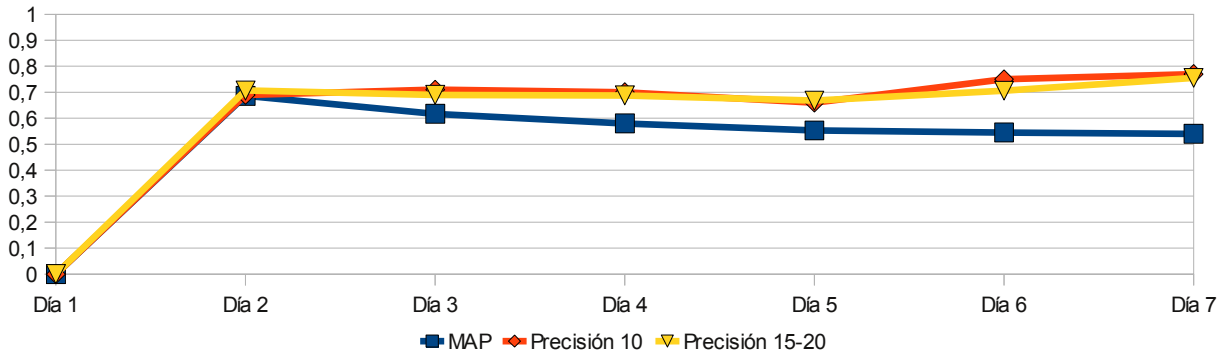

Figura 4.2. Progreso general del sistema durante una semana con el enfoque de pesos variables

El enfoque de tomar pesos uniformes para toda la noticia parece que hace que el perfil aprenda más despacio que en los enfoques anteriores respecto a las medidas de la precisión, según aparece en la figura 4.3, ya que en casi toda la semana tiende a empeorar, menos los dos últimos días, que se produce un aumento para acabar con valores similares de los enfoques dos enfoques anteriores de pesos variables. Hay que destacar que la medida de MAP disminuye en menor medida llegando a mantenerse constante antes que en los casos anteriores.

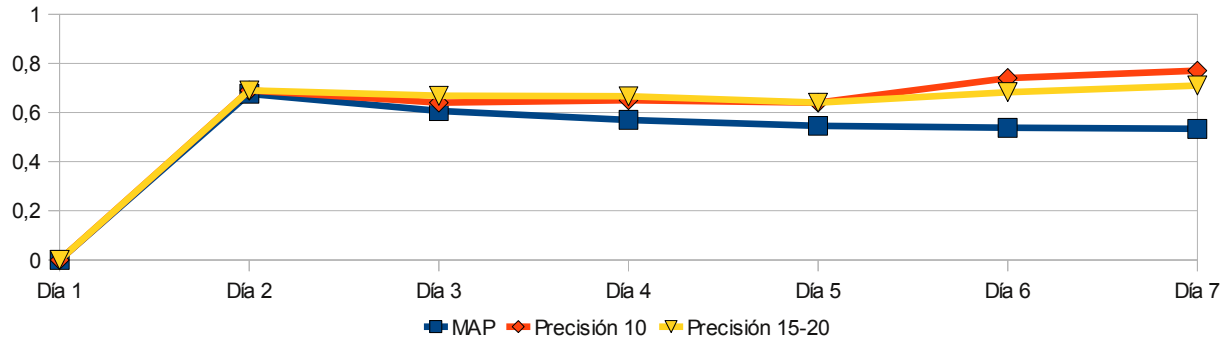

Figura 4.3. Progreso general del sistema durante una semana con el enfoque de pesos uniformes

Resultados casi idénticos son los que se obtienen con el enfoque que sólo toma el contenido de la noticia, ignorando el titular y el resumen, que aparece en la figura 4.4. En este enfoque se puede apreciar que el perfil incluso aprende más lentamente con respecto a la precisión, al incorporarse más ruido al perfil de los usuarios, ya que los términos más significativos, los que llaman la atención realmente a los usuarios, se encuentran tanto en el titular como en el resumen.

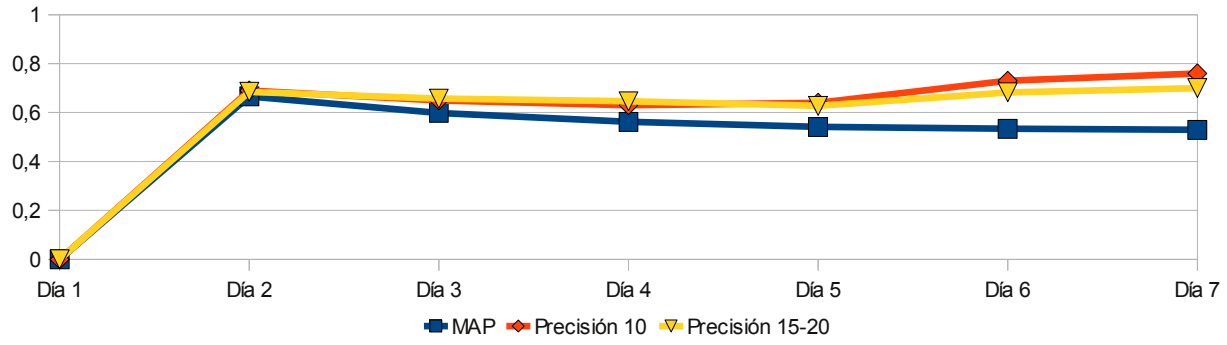

Figura 4.4. Progreso general del sistema durante una semana con el enfoque de tomar sólo el contenido

Desde un punto de vista global, las métricas de precisión tienden a aumentar a medida que pasa el tiempo, hasta obtener valores similares, al final de la semana. Lo que diferencia un enfoque de otro es la velocidad de aprendizaje, estando claro que se aprende de manera más rápida con el enfoque que toma pesos variables con el titular y el resumen. Por otro lado, la métrica de MAP tiende a disminuir en los cuatro enfoques, en menor o mayor medida, para terminar tendiendo a ser constante los últimos días de la evaluación.

4.2 Resultados con los grupos de usuarios definidos

Una vez deducido el enfoque con mejores resultados se sigue con la experimentación de los grupos definidos en el planteamiento. En la siguiente tabla se muestran los resultados utilizando el mejor enfoque respecto a la precisión sobre 10 y entre 15 y 20 documentos, es decir tomando únicamente el titular y el resumen de las noticias. Para cada grupo se muestran las medias de MAP, P10 y P15-20, durante el transcurso de la semana

	Interés Bajo			**Interés Medio**			**Interés Alto**		
	MAP	**P10**	**P15-20**	**MAP**	**P10**	**P15-20**	**MAP**	**P10**	**P15-20**
Día 1	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Día 2	0,6554	0,7000	0,6200	0,7092	0,7500	0,7720	0,7953	0,7000	0,7300
Día 3	0,5114	0,6750	0,6312	0,6813	0,8000	0,7815	0,7574	0,8500	0,7850
Día 4	0,4369	0,6250	0,5775	0,6353	0,7750	0,7146	0,7599	0,8500	0,8150
Día 5	0,4023	0,6000	0,5857	0,5905	0,7750	0,7146	0,7599	0,8000	0,8150
Día 6	0,3852	0,6500	0,5952	0,5777	0,9000	0,7866	0,7412	0,8500	0,8415
Día 7	0,3838	0,7000	0,6182	0,5718	0,8500	0,8070	0,7387	0,8500	0,8700
Media	**0,4625**	**0,6583**	**0,6046**	**0,6276**	**0,8083**	**0,7627**	**0,7587**	**0,8167**	**0,8094**

Tabla 4.4. Resultados con las métricas MAP, P10 y P15-20 para los diferentes grupos de usuarios

Si tomamos las medias de los MAPs de los tres grupos, como se muestra en la tabla, se deduce que el grupos con interés alto destaca sobre los de intereses medio y bajo.

Comparación	**Porcentaje de mejora**
In. Alto > In. Medio	20,88%
In. Alto > In. Bajo	64,04%

Tabla 4.5. Comparación usando la métrica MAP entre los tres grupos de usuarios

Como es muestra en la figura, el valor de MAP con respecto al tiempo para usuarios a los que tiene un interés bajo en las noticias tiende a disminuir bastante rápido. Sin embargo, los grupos de interés medio y alto, sobre todo este último, sufren una ligera disminución, hasta estabilizarse.

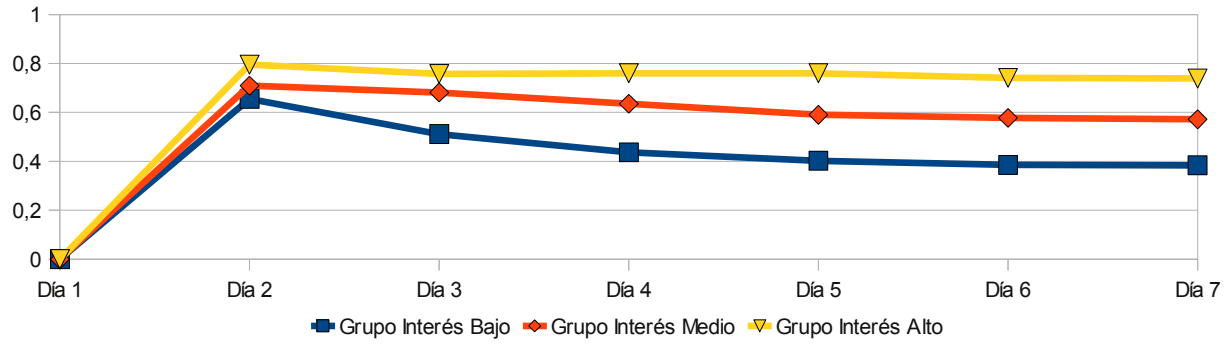

Figura 4.5. Comparación de la métrica MAP para los diferentes grupos de usuarios.

En el caso de tomar los valores de la métrica P10, la diferencia entre los grupos de interés alto y bajo son también significativas. Sin embargo, a la hora de comparar la precisión entre los grupos de interés medio y alto, se han deducido prácticamente los mismo valores con una mejora que apenas llega a una centésima parte.

Comparación	Porcentaje de mejora
In. Alto > In. Medio	0,96%
In. Alto ~ In. Bajo	24,06%

Tabla 4.6. Comparando usando la métrica P10 entre los tres grupos de usuarios

En la siguiente figura 4.6 se muestra cómo los grupos de interés medio y alto tiene prácticamente la misma progresión durante la semana, sin embargo, el grupo de interés bajo se aleja de los anteriores, obteniendo valores menores de precisión.

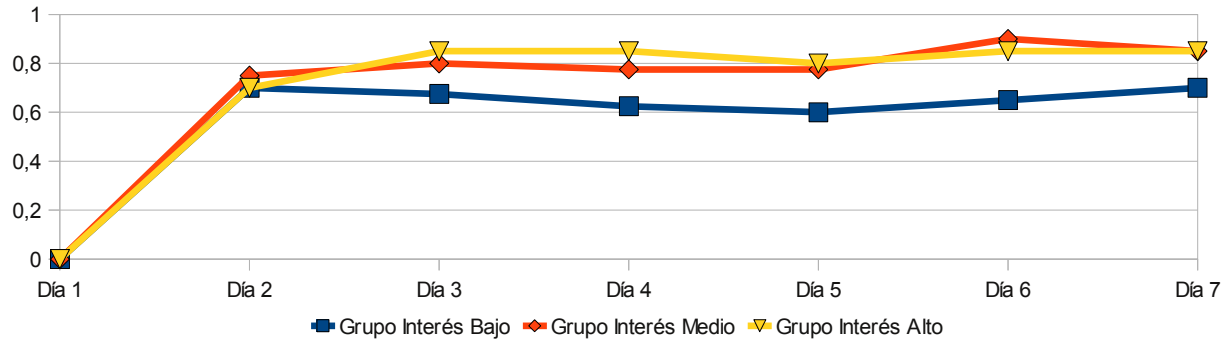

Figura 4.6. Comparación de la métrica P10 para los diferentes grupos de usuarios.

Como ha ido pasando en los experimentos anteriores, pero ahora en el caso de tomar los valores de la métrica P15-20, la diferencia entre los grupos de interés alto y bajo son también significativas. Por otra parte, si se compara con el grupo de interés medio, se puede observar que, a la hora de hacer un mayor número de recomendaciones se obtienen resultados ligeramente diferentes. Sin embargo, si bien el grupo de interés bajo tiende a mantenerse en esta precisión en torno al 60%, los otros dos grupos tienen a mejorar esta precisión a lo largo del tiempo y tienden a converger al mismo valor.

Comparación	Porcentaje de mejora
In. Alto > In. Medio	6,12%
In. Alto > In. Bajo	33,87%

Tabla 4.7. Comparando usando la métrica P10 entre los tres grupos de usuarios

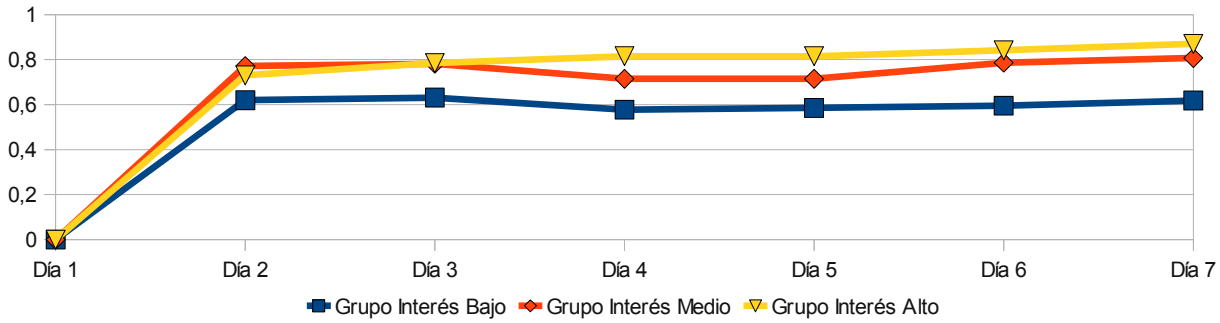

Figura 4.7. Comparación de la métrica P15-20 para los diferentes grupos de usuarios.

V. CONCLUSIONES

1 Introducción

Retomando las distintas fases del objetivo expuesto en el capítulo de introducción y analizando todo el trabajo realizado, se puede llegar a la conclusión más general de que se ha logrado alcanzar todas las metas establecidas.

En primer lugar, el tipo de información a personalizar, como la información periodística diaria. A la hora de encontrar la fuente, en el segundo objetivo de esta investigación, se optó por el proveedor EuropaPress, por ser un importante proveedor de noticias tanto para otras empresas de la información como para usuarios individuales. Elegir el formato más adecuado para cada elemento de información, que consiste en cada noticia, llevó a dividir la noticia en tres bloques, titular, resumen y contenido, para combinarlos de alguna manera de forma que se obtuvieran buenos resultados y proporcionándoles unos pesos específicos, ya sea omitiendo algunos de los bloques en el análisis y el proceso de similitud o bien dando más importancia a unos que a otros.

Para llevar a cabo la tercera tarea sobre la representación del perfil del usuario se tuvo que ver los ventajas y los inconvenientes de varios modelos, optando por un modelo mixto que combina estadística y estructura de datos con prioridad. A parte, el aprendizaje del mismo se realiza siguiendo la fórmula de Rocchio con algunas modificaciones para este trabajo de investigación. También ha sido necesario añadir un factor de olvido para que este perfil se fuera adaptando diariamente a los gustos e intereses del usuario

Una de las tareas más importantes es la cuarta, la manera de realizar la similitud entre los elementos de información, es decir, cada una de las noticias con el perfil de un usuario. Cuando se consiguió dar con las representaciones de los contenidos y del perfil del usuario, el modelo de espacio vectorial fue la manera de determinar qué posición tendría una noticia en un rankin según el perfil de ese usuario en concreto.

Finalmente, el ajuste de parámetros, donde se fijan los pesos de los diferentes bloques de una noticia, los analizadores para tratar y analizar la información de las noticias y la frecuencia de consulta a la fuente de contenidos para obtener las noticias se establecieron en la etapa de evaluación.

En este último capítulo del trabajo se van a describir las conclusiones que se pueden llegar con el uso de usuarios ideales, tanto de tipo equilibrado como focalizado, y ello permitió discernir, en un principio, los parámetros de los pesos de los bloques de las noticias para la segunda etapa de la evaluación, donde entran en juego usuarios reales, de los que se describen las conclusiones generales y en el caso de dividir a los usuarios en grupos. Por último, se cuenta el trabaja futuro que se pretende seguir, que consiste fundamentalmente en añadir un tratamiento más refinado de los contenidos y su representación, ya sea mediante entidades nombradas o conceptos, además, tampoco se descarta el remodelado del perfil del usuario para intentar hacerlo más general.

2 Conclusiones de los usuarios ideales

La utilización de los pesos de forma variable en cada uno de los bloques de la noticia ha mejorado los resultados, como se ve en el segundo y en el tercer enfoque, aunque en este último, el recall que se obtiene es demasiado bajo, debido al aumento de la precisión.

También se pueden deducir de estos resultados que existe una relevante diferencia entre un perfil focalizado y uno equilibrado, obteniendo mayor precisión en el primero. A pesar de la probabilidad a la hora de preferir una categoría, las noticias que se proponen están dentro de unos márgenes de proporción para cada perfil.

3 Conclusiones de los resultados generales

En los resultados generales obtenidos de la evaluación con usuarios reales se puede concluir de forma inmediata que, si se persigue una precisión elevada en un dispositivo móvil, en el que tiene sentido proporcionar entre diez y veinte noticias recomendadas, la mejor configuración es tomar los bloques de titular y resumen de una noticia con pesos variables. La diferencia entre el resto de enfoques y el de tomar estos dos bloques es significativa.

Por otro lado, si lo que se pretende es tener la mejor media de recomendaciones, entre los enfoques de tomar pesos variables de toda la noticia, o bien, de los bloques titular y resumen, cualquiera de los dos es una buena opción ya que la diferencia no es casi significativa. Los otros enfoques se quedan bastante atrás en este ámbito por añadir ruido en el perfil del usuario y a la hora de aplicar el factor de olvido no se consigue converger a tiempo con los gustos del usuario.

Si se tiene que optar por algún enfoque de forma global, sin duda, el enfoque de pesos variables tomando sólo los dos bloques de titular y resumen es la mejor opción para la personalización en un entorno móvil.

4 Conclusiones de los grupos de usuarios

Si en lugar de ver a todos los usuarios de forma general, se dividen por el nivel de noticias que les interesan, estableciendo los grupos de interés alto, medio y bajo, los resultados para los dos primeros son significativamente buenos y muy parecidos, en cambio, para el grupo de interés bajo, los resultados distan bastante de los otros. Los parámetros se establecen según los mejores resultados que se han obtenido en los generales, es decir, para estos experimentos se tiene un enfoque de pesos variables tomando los bloques del titular y el resumen.

Tanto para el grupo de interés alto y medio la precisión es más del 80%, sin embargo, en el caso del grupo de interés bajo apenas alcanza el 66%. De la misma forma, para la métrica MAP, aunque la diferencia entre el grupo de interés alto e interés medio se acentúa, la diferencia entre estos dos grupos, con respecto al grupo de interés bajo sigue siendo importante.

La gran diferencia entre los grupos es debida, principalmente, a la falta de información que sufren los perfiles de los usuarios de estos grupos, por lo tanto habría que investigar otro tipo de representación del perfil para los usuarios de este grupo. También se puede deducir que el grupo de interés bajo es como una 'carga' para los resultados generales, ya que hace caer la media de manera importante.

5 Trabajo futuro

Después de todo el análisis de los datos y resultados por parte de los usuarios ideales y por los usuarios reales, tomando varios enfoques a la hora de representar una noticia y los posibles pesos de cada parte de la noticia, se puede decir que se puede insertar más procesamiento sobre las noticias para mejorar su representación y de forma directa la del perfil del usuario.

Este procesamiento añadido consiste en el tratamiento de entidades nombradas y el uso de conceptos en vez de palabras o sus raíces. Las entidades nombradas pueden proporcionar información más refinada para diferencias las diferentes temáticas de una cierta categoría, incluso, influir de alguna manera para rechazar ciertas noticias si esta entidad se toma como algo 'negativo' para el usuario. El uso de conceptos y de esta manera representar las noticias y el perfil del usuario con ellos, también ayuda a poder definir mejor los intereses que palabras o raíces independientes que pueden ser ambiguas en varios temas de una misma categoría.

Por otro lado, el uso de varias fuentes de contenido, ya no sólo del ámbito periodístico, si no foros, restauración, etc... Para probar y refinar el modelo del usuario con el fin de que se pueda aplicar en cualquier sistema.

Por último, aprovechando que la investigación se realiza sobre dispositivos móviles y su tecnología avanza a pasos agigantados, se puede aprovechar el componente de geo-localización para realizar una personalización dependiendo en el punto geográfico que se encuentra en relación con en el momento del día.

BIBLIOGRAFÍA

[1] Al Masum, S.M; Islam, M.T.; Ishizuka M. (2006) ASNA: *An Intelligent Agent for Retrieving and Classifying News on the Basis of Emotion-Affinity*. Proceedings of the International Conference on Computational Inteligence for Modelling Control and Automation and International Conference on Intelligent Agents Web Technologies and International Commerce, pp. 6.

[2] Albayrak, S; Wollny, S; Varone, N; Lommatzsch, A; Milosevic, D. (2005) *Agent technology for personalized information filtering: the PIA-system*. Proceedings of the 2005 ACM symposium on Applied computin, pp. 54-59.

[3] Benyon, D. (1994). "Accommodating Individual Differences through an Adaptive User Interface". Presented by Alison Nichols, October 21 st 1994.

[4] Berghel, H. (1997). *Cyberspace 2000: Dealing with information over-load*. Communications of the ACM 40:19–24.

[5] Billsus, D; Pazzani, M.J. (2007) *Adaptive news access. Lecture Notes in Computer Science* 4321, pp. 550-570.

[6] Bowman, C.M.;Danzig, P.B.; Manber, U.;Schwartz M.F. Scalable internet resource discovery: Research problems and approaches. CACM, 37(8):98{107, August 1994.

[7] Chan, C.C.H. (2008) *Intelligent spider for information retrieval to support mining-based price prediction for online auctioning*. Expert Systems with Applications 34(1), pp. 347-356.

[8] da Cruz R.; García, F.; Romero, L. (2003) *Perfiles de usuario en la senda de la personalización*. Technical report, Universidad de Salamanca – Departamento de Informática y Telemática, Enero 2003.

[9] De Bra, P; Aerts, A; Houben, G.J.; Wu, H.(2000) *Making General-Purpose Adaptive Hypermedia Work.* Proceedings of the WebNet Conference. Pp.117-123.

[10] Deagostini, A; Cormenzana, F. (2005). *Interfaces de usuario Inteligentes: Sistemas Adaptativos*. Interacción humano-computador y diseño de interfaces.

[11] Fellbaum C. (1999) *WordNet: An Electronic Lexical Databases*, MIT Press, Cambridge, Massachusetts, 1999.

[12] Froufe, A; Jorge, P. (2004) *J2ME Java 2 Micro Edition Manual de Usuario y tutorial*. Ra-Ma

[13] García-Cabrera, L; Rodríguez-Fortiz, M.J.; Parets-Llorca, J. (2001). *Formal Foundations for the Evolution of Hypermedia Systems*. 5th European Conference

on software Maintenance and reengineering, Workshop on FFSE. IEEE Press. March 5-12. Lisboa, Portugal.

[14] García-Cabrera, L. (2001). *SEM-HP: A Systemic, Evolutionary, Semantic Model for Hypermedia System Development. (in Spanish)*. Ph Thesis. November 2001.

[15] Gauch, S; Speretta, M; Chandramouli, A; Micarelli, A.(2007) *User profiles for personalized information access*. Lecture Notes in Computer Science 44321, pp. 54-89.

[16] Google Webmasters (2010). Cómo saber que el robot es Googlebot [en línea]. Último acceso el 16 de abril de 2010. http://www.google.com/webmasters/

[17] Gövert, N.; Lalmas, M.;Fuhr, N. (1999). *A probabillistic description-oriented approach for categorising Web documents*. In Proceedings of CIKM-99, 8th ACM International Conference on Information and Knowledge Management (Kansas City, MO, 1999), 475–482.

[18] N. Henze and W. Nejdl. (1999) *Bayesian modeling for adaptive hypermedia systems*. In ABIS 99, 7. GI-Workshop Adaptivitat und Benutzermodellierung in interaktiven Softwaresystemen, Magdeburg, Sept. 1999.

[19] HR-XML Consortium (2010). *The independent platform for development of human resources XML vocabularies* [en linea]. Ultimo acceso 16 de Abril de 2010. http://www.hr-xml.org/

[20] IMS Global Learning Consortium (2010). *IMS Learner Information Package Specification* [en linea]. Ultimo acceso 16 de Abril de 2010. http://www.imsglobal.org/profiles/

[21] Indulska, J; Robinson, R; Rakotonirainy, A; Henricksen. (2003) K. Experiences in *Using CC/PP in Context-Aware Systems*. En Chen, M.-S.,Chrysance, P.K., Sloman, M. Zaslavsky, A. (Eds.),Proceedings of the 4th International Conference on Mobile Data Management NCS, Springer-Verlag, Vol. 2574, pp247-261.

[22] Lee, C.S.; Chen, Y.J.; Jian, Z.W. (2003) *Ontology-based fuzzy event extraction agent for Chinese e-news summarization*. Expert Systems with Applications 25(3), pp. 431-447.

[23] Lee, W.P.; Yang, T.H. (2003) *Personalizing information appliances: a multi-agent framework for TV programme recommendations*. Expert Systems with Applications 25(3), pp. 331-341.

[24] Liu, H.; Singh, P. (2004) *"ConceptNet: A Practical Commonsense Reasoning Toolkit"*, BT Technology Journal, Vol 22, No 4, Oct. 2004, pp. 211-226.

[25] Marín, D; Rico, A (2009). *Modelo para la adaptación de información en ambientes nómadas*. Revista Sistemas. pp. 69-80

[26] Methanol Web Crawling System (2010). Methabot Documentation [en línea].

Último acceso el 16 de abril de 2010. http://metha-sys.org/

[27] Mobile User Experience. (2010) Manifesto [en línea]. Último acceso el 19 de Abril de 2010. http://www.pmn.co.uk/mex/

[28] Medina-Medina, N; García-Cabrera, L; Rodríguez-Fortiz, M.J.; Parets-Llorca, J. (2001).Adaptación al Usuario en Sistemas Hipermedia: El Modelo SEM-HP.

[29] Nakashima, T.; Nakamura, R. (1997). *"Information Filtering for the Newspaper"*. IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, August 1997. Victoria, B.C., Canada

[30] Pazzani, M.m; Billsus, D. (2007) Content-based Recommendation Systems. In: Brusilovsky, P., Kobsa, A., Nejdl, W. The Adaptative Web: Methods and Strategies of Web Personalization. Lecture Notes in Computer Science, Vol 4321

[31] Rich, E. (1979) *Building and Exploiting User Models*. in Department of Computer Science.Pittsburgh, PA: Carnegie-Mellon University.

[32] Rich, E. (1979) *User Modeling via Stereotypes*. Cognitive Science 3, 329-354.

[33] Rich, E. (1983) *Users are Individuals: Individualizing User Models*. International Journal of Man-Machine Studies 18, 199-214.

[34] Rocchio, J.J. Jr., 1971. "*Relevance feedback in information retrieval"*, The SMART Retrieval System: Experiments in Automatic Document Processing, Prentice-Hall.

[35] Sarawagi, S. (2008) *Information extraction. Foundations and Trends in Databases* 1(3), pp. 261-377.

[36] Sebastiani, F., 1999. *"A tutorial on automated text categorization"*. Proceedings of ASAI-99, 1st Argentinian Symposium on Artificial Intelligence, pp. 7-35.

[37] Sobecki, J.; Szczepanski, L. (2007) *Wiki-News Interface Agent Based on AIS Methods*. Lecture Notes in Computer Science 4496, pp. 258-266.

[38] Tu, H.C.; Hsiang, J. (2000) *An architecture and category knowledge for intelligent information retrieval agents*. Decision Support Systems 28(3), pp. 255-268.

[39] Yahoo! Help (2010). Yahoo! Web Crawler Information [en línea]. Último acceso el 16 de abril de 2010. http://help.yahoo.com/l/us/yahoo/search/webcrawler/

[40] Yang, Y. (1999) . *"An evaluation of statistical approaches to text categorization".* Information Retrieval, Vol. 1, Number 1-2, pp. 69-90.

[41] Yu, H.; Mine, T.; Amamiya, M. (2009) *Agent-Community-based P2P semantic MyPortal information retrieval system architecture*. Journal of Embedded Computing 3(1), pp. 63-75

[42] Wu, H.; Houben, G.J.; De Bra, P.; *"Supporting User Adaptation in Adaptive Hypermedia Applications"*, Proceedings InfWet2000. Rotterdam, the Netherlands.

ANEXO I. RECOPILACIÓN DE INFORMACIÓN

1 Introducción

A pesar de idear un sistema de personalización completamente implícito se requieren datos explícitos para probar si realmente el sistema se adapta de manera adecuada a lo largo del tiempo.

Esta recopilación de información de los usuarios se ha llevado a cabo durante los siete días de la semana entre el 12 de Mayo de 2010 y el 19 de Mayo de 2010, con las noticias obtenidas a las 9:30 de la mañana.

El procedimiento consistía en que el sistema, una vez que obtenía las noticias de la fuente de contenidos, creara un formulario, que consistía en un documento Excel con un formato específico, que se lo envía a un conjunto determinado de usuarios, los cuales lo reenviaban ya resuelto.

El formato y una instancia de estos formularios de obtención de datos de relevancia de los usuarios se explican en los apartados siguientes de este anexo.

2 Formato del formulario de obtención de datos de los usuarios

La evaluación del sistema ha requerido un seguimiento explícito de los usuarios para saber hasta que punto las salidas obtenidas eran buenas. Para ello, se ideó un formato específico de formulario, el cual se enviaba todos los días a cada usuario. Debía de ser sencillo y relativamente rápido de realizar ya que al día se obtenían al rededor del centenar de noticias.

Este formato es el que se muestra en la tabla 2.1. Consta de tres columnas, la primera es donde marca el usuario con un 1 o un 0, la segunda es la noticia y la tercera es el identificador de la noticia en la base de datos para automatizar la parte de evaluación.

Interesa (1) No Interesa (0)	**Titular**	**ID Noticia**
	Titular 1	id1
	….	….
	Titular n	idn

Tabla 2.1. Formato de las encuestas para los usuarios.

3 Ejemplo de instancia para el primer día

La siguiente tabla es la que se usa para el primer entrenamiento del sistema, y es la que en primer lugar se enviaba a los usuarios para obtener la precisión del sistema. Las noticias se han obtenido de EuropaPress el doce de mayo de 2010.

A parte de este, un documento Excel con 104 noticias, existen seis documentos más de los seis días siguientes que se enviaban periódicamente a los usuarios. Se pueden encontrar en la carpeta *datosEvaluación* del desarrollo del proyecto.

Interesa	Titular	ID
	El Atlético quiere ser campeón de la Europa League	13298
	La vida de Deker peligra en Los hombres de Paco	13299
	Arnold Schwarzenegger, un héroe contra El fin de los días	13300
	Numb3rs despide su sexta temporada con boda	13301
	TVE rinde homenaje a Antonio Vega	13302
	LaSexta adelanta la cobertura de la Fórmula 1 en Montecarlo	13303
	El PP acusa a TVE de "ocultar" a Rajoy	13304
	Si no hay más demanda no habrá TDT en movilidad	13305
	Trancas y Barrancas ya tiene móvil	13306
	Fantasmas y pescadores en Alaska, los realities más sorprendentes	13307
	Amaia Salamanca: Yo no me veo parecida a Letizia	13308
	Evangeline Lilly: "El final de Perdidos será fiel al estilo de la serie"	13309
	Quique: "La final da sentido a todo lo hecho hasta ahora"	13310
	Hodgson: "No queremos ser segundos, estamos convencidos de terminar llevándonos un trofeo a casa"	13311
	La gloria se encuentra en Hamburgo	13312
	Simao: "La final está al cincuenta por ciento, pero quiero ganarla"	13313
	Palmarés de la competición	13314
	Villa: "No quiero vivir lo del año pasado"	13315
	Rosell: "Fichar ahora a Villa costaría el doble que después de las elecciones"	13316
	Rosell: "Queremos que Guardiola sea nuestro Ferguson"	13317
	Sandro Rosell, cantera, cracks y Guardiola	13318
	La ACB garantiza la disputa de la competición	13319
	Alonso quiere seguir a ritmo de podio	13320
	El Gobierno baja un 5% el sueldo de los funcionarios desde el verano	13321
	El Gobierno no descarta nuevas medidas fiscales	13322
	Zapatero: las medidas son "imprescindibles"	13323
	Zapatero defiende que los que tienen más ingresos hagan mayor esfuerzo	13324
	Rajoy pide a Zapatero suprimir vicepresidencia tercera	13325
	Rajoy no apoyará congelar pensiones si no recorta subvenciones a partidos y sindicatos	13326
	Las minorías ven improvisado el recorte drástico del gasto social	13327
	El TS ordena al TSJV seguir investigando a Camps por recibir trajes	13328
	El PPCV respaldará esta tarde a Camps en una Junta Directiva	13329
	El Supremo abre juicio oral contra Garzón	13330
	Varela: "la protección de las victimas" no justifica su "irresponsabilidad"	13331
	Zapatero: cuanto más se hable del TC, más difícil será su renovación	13332
	Cameron y Clegg ultiman la composición del Gobierno	13333
	El nuevo ministro de Exteriores dice que Afganistán seguirá siendo una prioridad y rechaza ceder más poder a UE	13334
	Cameron llama a Sarkozy y se comprometen a "trabajar estrechamente"	13335
	Rusia expresa su deseo de mejorar sus relaciones con Reino Unido cuando se forme el nuevo gobierno	13336

	Un niño holandés de diez años sobrevive a un accidente de avión en Libia	13337
	Un total de 61 holandeses, en el avión accidentado en Libia	13338
	Airbus promete asistencia técnica a las autoridades para determinar las causas del accidente aéreo en Trípoli	13339
	Afriqiyah expresa sus condolencias por las víctimas del siniestro y rechaza "especular" sobre las causas	13340
	Siete niños muertos apuñalados en una guardería del noroeste de China	13341
	Investigan los presuntos ataques con gas a escuelas de niñas en Afganistán	13342
	El Papa se salta las medidas de seguridad para saludar a los niños	13343
	Ahmadineyad: "Las resoluciones de la ONU no valen un centavo"	13344
	El recorte de sueldos públicos permitirá ahorrar al menos 2.400 millones	13345
	El Gobierno cifra en 670 millones el ahorro en dependencia y en 300 el de gasto farmacéutico	13346
	Almunia califica las medidas del Gobierno de "paso lógico"	13347
	De la Vega explicó a sindicatos de función pública el recorte	13348
	La obra pública sufrirá retrasos medios de entre seis meses y un año	13349
	El Santander dice que las medidas de Zapatero "van en la buena dirección"	13350
	Méndez advierte de que las nuevas medidas de ajuste van a desatar el conflicto	13351
	Toxo advierte que las medidas de ajuste merecen una contestación "masiva"	13352
	El sindicato de funcionarios no descarta una huelga general	13353
	Bruselas propone más sanciones por incumplir el Pacto de Estabilidad	13354
	Zapatero quiere acelerar la reforma de cajas	13355
	Trichet achaca los problemas actuales a las políticas de cada país	13356
	La Audiencia Nacional rechaza la demanda de los controladores	13357
	Miembros de la CEOE piden a Díaz Ferrán que solucione sus problemas	13358
	Richard Serra, el triunfo del minimalismo "audaz"	13359
	Subastan la única grabación de una rueda de prensa de los Beatles	13360
	El Príncipe de Persia en acción	13361
	Fito Páez presenta en Madrid y Barcelona su nuevo disco	13362
	El Capitán América y Los Vengadores ¿en 3D?	13363
	El canon digital español es ilegal	13364
	Un nuevo primate complica la comprensión de la evolución	13365
	Los laboratorios de la NASA se quedan anticuados	13366
	Localizan un agujero negro supermasivo en retroceso	13367
	El primer detector de neutrones por centelleo tiene sello español	13368
	Las arañas inspiran una seda cinco veces más fuerte que el acero	13369
	Herschel descubre en primicia un agujero en el espacio	13370
	Todos los asteroides y cometas que acechan la Tierra, en una web	13371
	Dos europeos, tres rusos y un chino viajarán a Marte en simulador	13372
	El LHC sigue batiendo marcas operativas	13373
	Es cuestión de tiempo encontrar Pandoras	13374
	Aparecen en Ibiza restos de un muro romano del siglo IV d.C.	13375
	Descubren el cluster de galaxias más lejano del Universo	13376

	Ford inicia la producción en México del nuevo Fiesta	13377
	Aston Martin fabrica la primera unidad del nuevo Rapide	13378
	Bosch inicia la producción del sistema predictivo de frenada de emergencia	13379
	Telefónica y Endesa presentan la primera cabina telefónica que recarga coches eléctricos	13380
	Ford Almussafes fabricará sus primeros híbridos e híbridos enchufables para el mercado europeo	13381
	Sebastián ve en el coche eléctrico una apuesta por el futuro industrial, energético y medioambiental	13382
	Banesto Agrícola financiará los vehículos todoterreno de Tata Motors	13383
	Fiat lanzará después de verano el nuevo Fiat 500C by Diesel	13384
	Hertz e Iberia ofrecen descuentos especiales en los alquileres de vehículos	13385
	Seat regalará un curso a los clientes que compren sus modelos más deportivos	13386
	EE.UU abre una investigación a Toyota por una campaña de revisión de automóviles	13387
	Infiniti incorpora un motor diésel de 238 caballos a su crossover FX	13388
	Cuba solicitará el hábeas corpus para uno de los 5 cubanos detenidos en EEUU	13389
	BP tratará de sellar la fuga de petróleo con una campana metálica de menor tamaño	13390
	Venezuela libera a tres colombianos acusados de espionaje	13391
	Amplían la detención al detenido en la Embajada de EEUU en Chile	13392
	Los Cuerpos de Paz de EEUU regresarán a Colombia tres décadas después	13393
	Intelectuales y artistas firman un manifiesto por la libertad en Cuba	13394
	Brasil, primer país en aportar dinero para la reconstrucción de Haití	13395
	La FIFA suspende a la Federación Salvadoreña	13396
	Maradona deja fuera a Cambiasso, Gago y Zanetti	13397
	Venezuela libera a tres de los ocho colombianos detenidos en marzo acusados de espionaje	13398
	Brasil, primer país en aportar dinero al fondo para la reconstrucción de Haití	13399
	La cifra de desaparecidos en Colombia asciende a 38.255 en los últimos tres años, según un informe	13400

ANEXO II. RESULTADOS DE LOS INDIVIDUOS

1 Introducción

En esta anexo se proporciona de forma individualizada todos los datos que se han obtenido de la progresión del aprendizaje a lo largo del tiempo de cada uno de los usuarios.

En este análisis se muestran los individuos separados por grupos según los parámetros establecidos en la evaluación de este trabajo, es decir, por grupos de interés alto, medio y bajo.

Para cada uno de los individuos se muestra su progresión de interés de cada día de las noticias que se les ha presentado para la evaluación en una gráfica de barras y a continuación se muestra para cada uno, la progresión de las métricas de MAP, P10 y P15-20 que se han explicado en el planteamiento de la evaluación de este trabajo. Esta segunda característica se hace utilizando el enfoque de tomar únicamente el titular y resumen de las noticias y tomar pesos variables para cada bloque.

2 Grupo de interés alto

2.1 Usuario Usuario 3

Este usuario posee el segundo porcentaje de interés más alto, como se muestra en la gráfica 2.1. En concreto, los momentos de la semana en los que accede a más noticias es al principio y al final de la semana, habiendo una disminución del interés por, posiblemente cambios de temáticas.

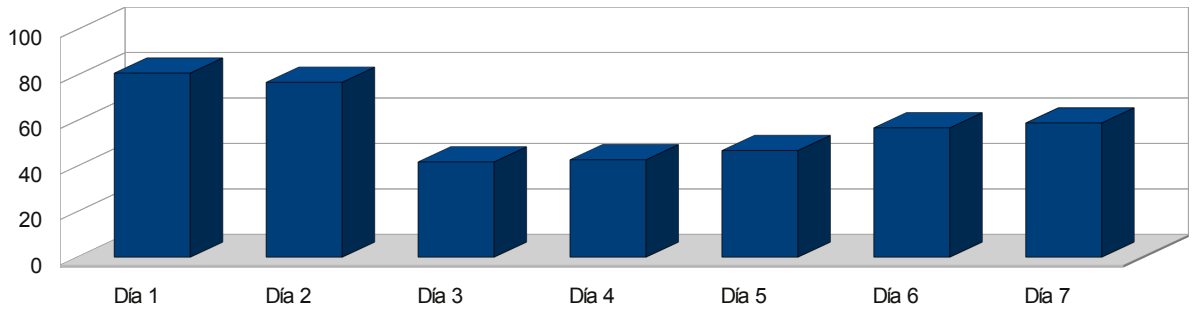

Figura 2.1. Número de noticias de interés por día del usuario Usuario 3

En la figura 2.2, se nota claramente la disminución de interés y el cambio de los temas de las noticias durante la semana, sin embargo, para los tres parámetros de MAP, P10 y P15-20, el la monotonía se mantiene constante, bien, creciente hasta el final de la semana.

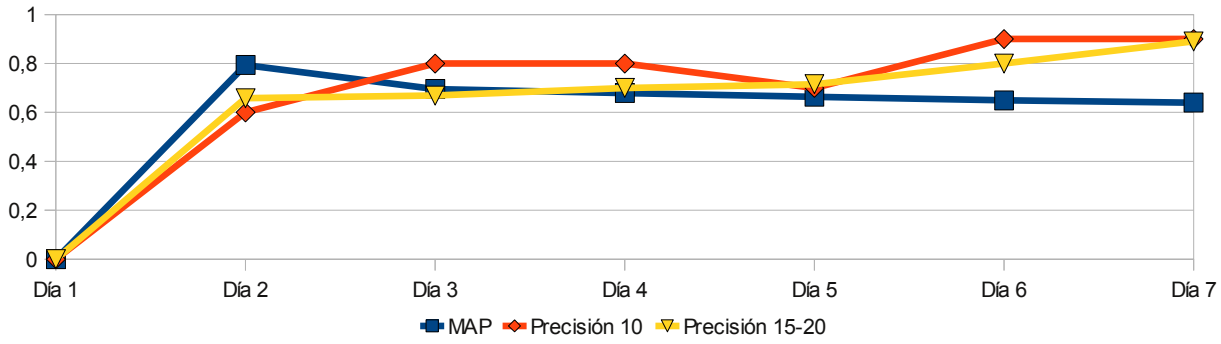

Figura 2.2. Progreso durante una semana del usuario Usuario 3

2.2 Usuario Usuario 7

El porcentaje de interés mas alto pertenece a este usuario. A pesar de que a principios de la semana es cuando menos interés tiene, éste va aumentando de forma progresiva durante toda la semana, es decir, aunque haya cambios sobre los temas durante la semana, eso agrada a este usuario, como se ve en la figura 2.3.

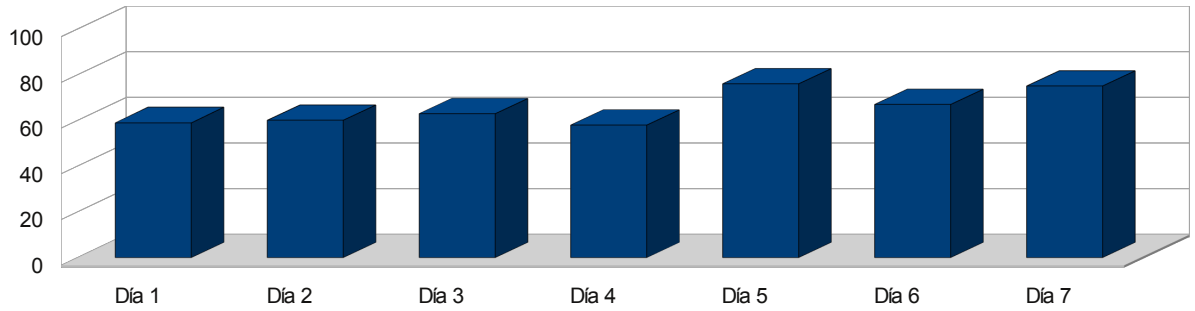

Figura 2.3. Número de noticias de interés por día del usuario Usuario 7

La progresión del aprendizaje que muestra la figura 2.4 denota que, desde el segundo día, las medidas que se utilizan, no bajan del 80%, manteniéndose constante durante toda la semana.

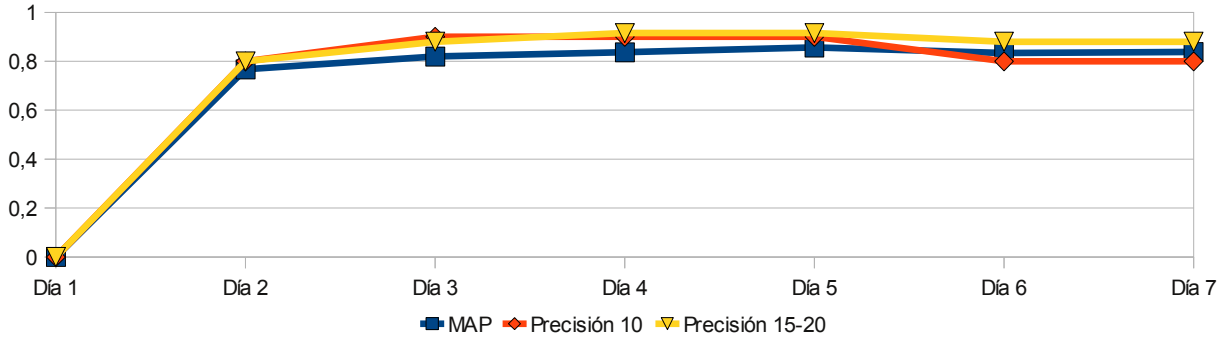

Figura 2.4. Progreso durante una semana del usuario Usuario 7

3 Grupo de interés medio

3.1 Usuario Usuario 4

La progresión de interés en este usuario, como se ve en la figura 3.1, es decreciente a lo largo de toda la semana, empezando con un interés del más del 50%, pero acabando apenas con un 20% de interés. Esto se atribuye a que los temas o acontecimientos que aparecían esa semana no eran de su interés.

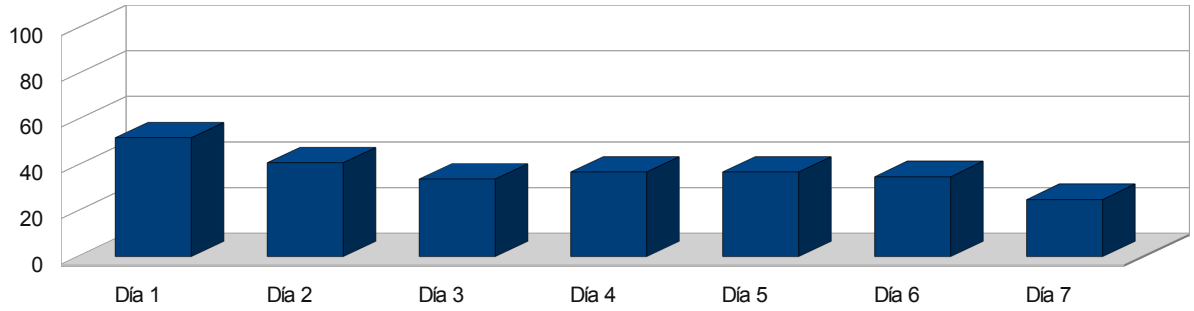

Figura 3.1. Número de noticias de interés por día del usuario Usuario 4

A pesar de que su interés disminuía a lo largo de la semana, en la figura 3.2 se pude ver que el sistema ha logrado mantener, con una disminución apenas apreciable, tanto de las precisiones como del MAP. En los dos días finales de la semana se produce un pico importante con respecto a la precisión sobre las 10 primeras noticias, mientras que las otras dos métricas se mantienen prácticamente constantes, con una disminución del MAP de casi un 10%. A pesar de que el interés es menor sobre con el conjunto de noticias, el sistema es capaz de recomendarle suficientemente bien.

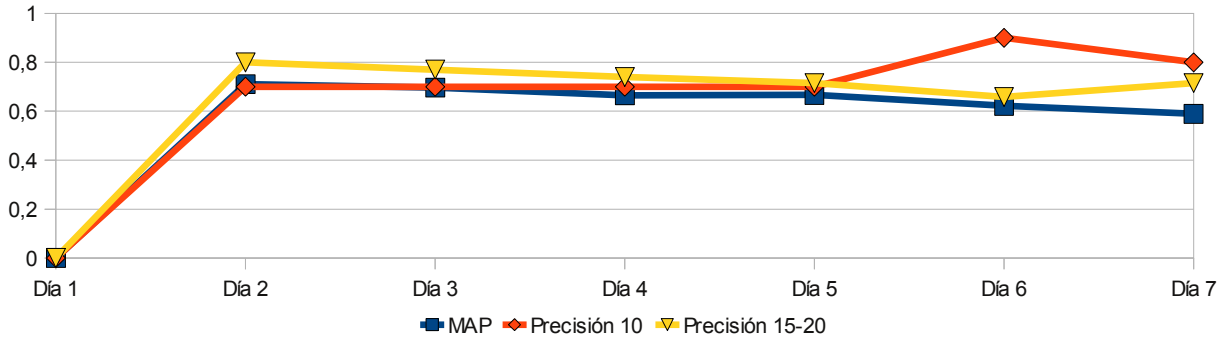

Figura 3.2. Progreso durante una semana del usuario Usuario 4

3.2 Usuario Usuario 2

Este usuario es típico de este grupo, es decir, como se muestra en la figura 3.3, su interés sobre el conjunto de las noticias varía cada día sin un patrón definido, como en el caso anterior que poseía una monotonía decreciente, habiendo picos de más del 70% y depresiones que apenas llegan al 25%.

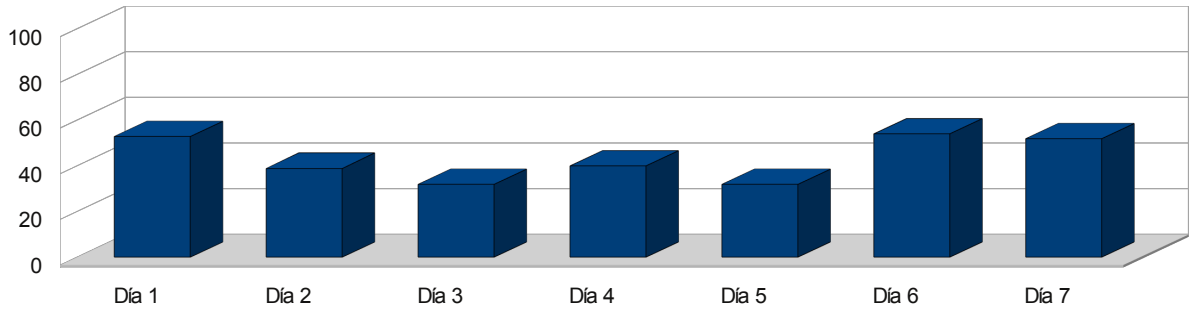

Figura 3.3. Número de noticias de interés por día del usuario Alvaro

Debido a la variabilidad de los intereses del usuario, como aparece en la figura 3.4, en las tres métricas ocurre el fenómeno de que el perfil del usuario tarda en aprender sus verdaderos intereses, los cuales, al final de la semana, se consiguen reflejar con una precisión de más del 90%.

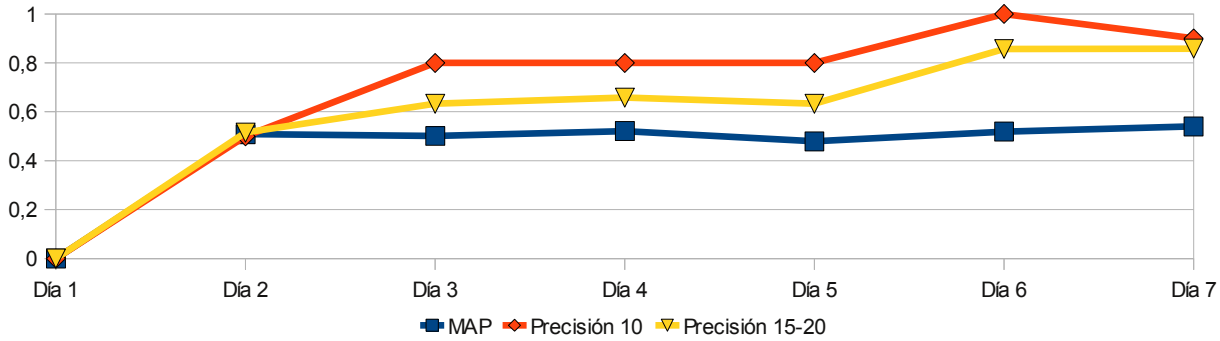

Figura 3.4. Progreso durante una semana del usuario Usuario 2

3.3 Usuario Usuario 5

Otro usuario típico de este grupo, aunque con picos y depresiones menos pronunciadas.

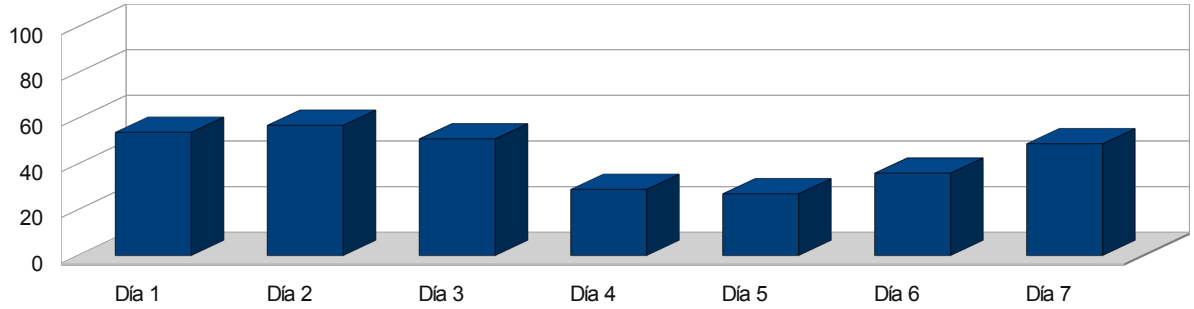

Figura 3.5. Número de noticias de interés por día del usuario Usuario 5

En la figura 3.6 se nota el efecto menos pronunciado de picos de interés, ya que desde los primeros días las métricas rozan el 90%. A pesar de ello, la métrica MAP empieza a disminuir hasta menos del 70%. Por otro lado, ambas medidas de precisión logran incrementarse hasta llegar a más del 95%.

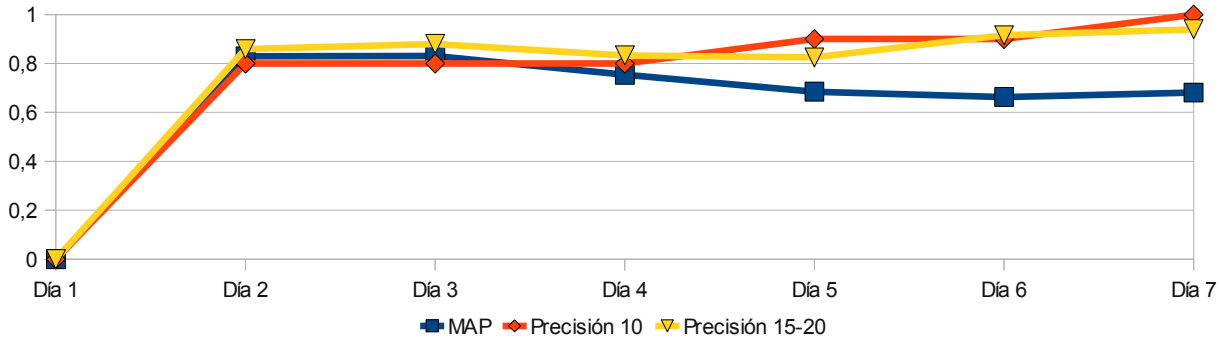

Figura 3.6. Progreso durante una semana del usuario Usuario 5

3.4 Usuario Usuario 9

Es el caso más pronunciado de variabilidad de intereses. Se tiene un pico de más del 50% y depresiones que no alcanzan el 10% de interés, como ocurre en el quinto día.

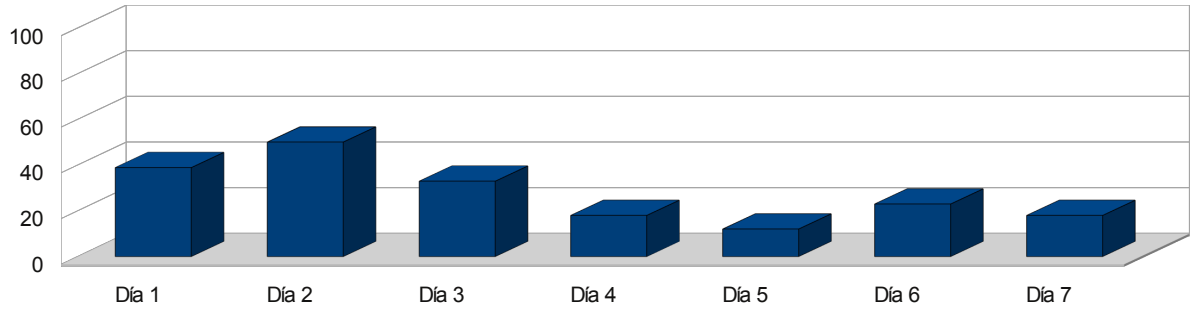

Figura 3.7. Número de noticias de interés por día del usuario Usuario 9

Los resultados de las métricas que aparecen en la figura 3.8 revelan que al principio de la semana se obtienen precisiones que alcanzan el 100% y un MAP de un poco más del 80%, pero debido a la disminución de interés a lo largo de los días estas medidas van disminuyendo o fluctuando, intentando ajustarse a los intereses de este usuario.

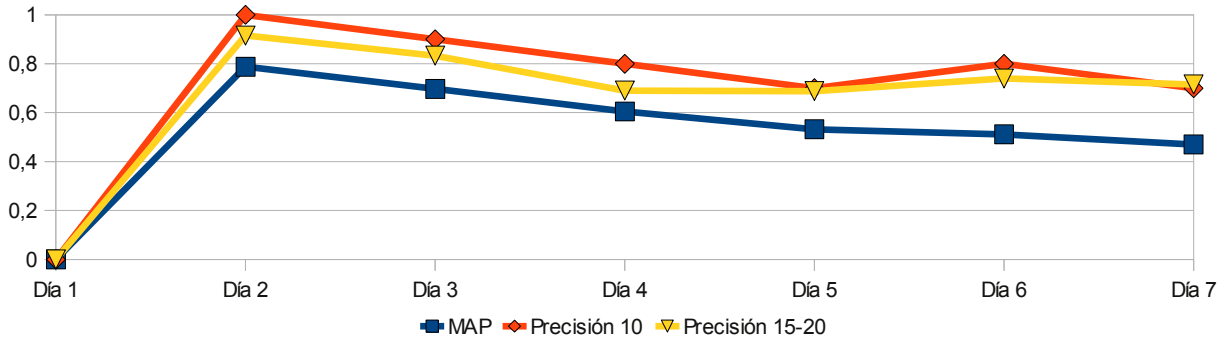

Figura 3.8. Progreso durante una semana del usuario Usuario 9

4 Grupo de interés bajo

4.1 Usuario Usuario 1

Este usuario, según la figura 4.1, a pesar de tener unos intereses diarios considerablemente constantes, en ninguno de los días se consigue superar el 20% de interés. No se va a tener demasiada información para ir definiendo su perfil.

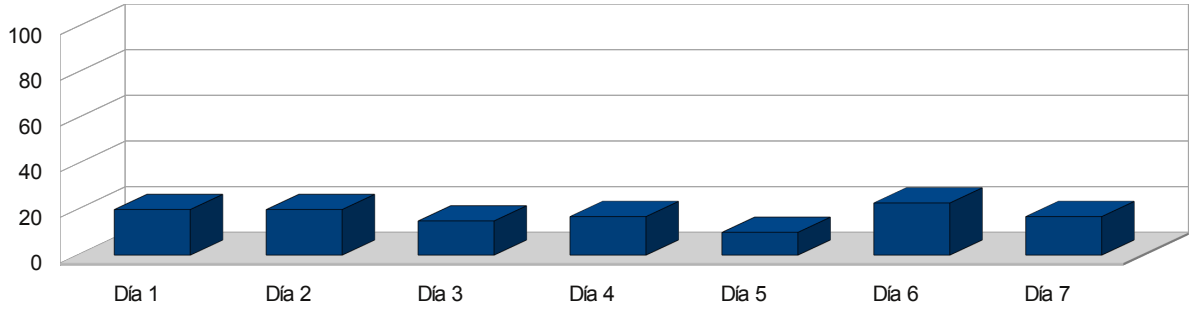

Figura 4.1. Número de noticias de interés por día del usuario Usuario 1

Como se ha comentado antes, con la poca información que consigue recopilar el sistema de manera implícita durante toda la semana, de la figura 4.2 se puede ver que le cuesta ajustar los intereses del usuario hasta obtener un 80% con respecto a la precisión de las 10 primeras noticias. Hay que destacar que el MAP se mantiene constante, de la misma forma que la precisión entre las 15 y 20 primeras noticias.

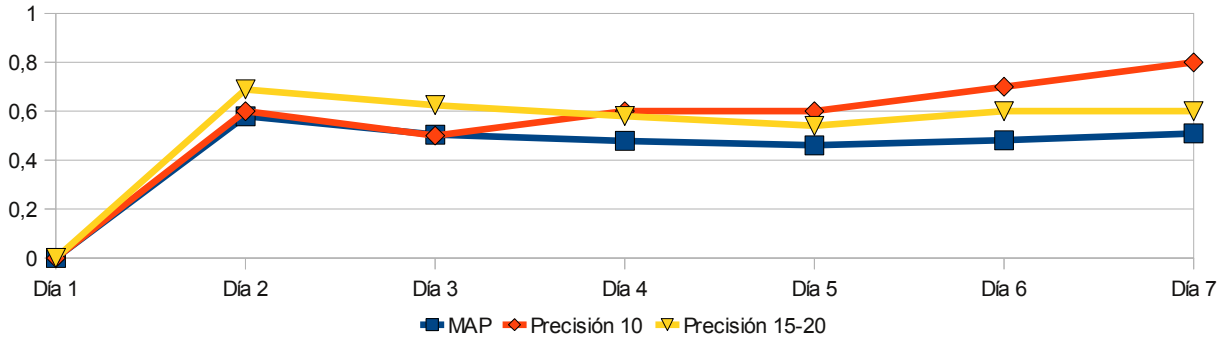

Figura 4.2. Progreso durante una semana del usuario Usuario 1

4.2 Usuario Usuario 6

Otro usuario con interés muy bajo sobre el conjunto de todas las noticias. El primer día supera el 20% de interés, sin embargo, hay tres días que apenas se llega al 10%, como apetece en la figura 4.3.

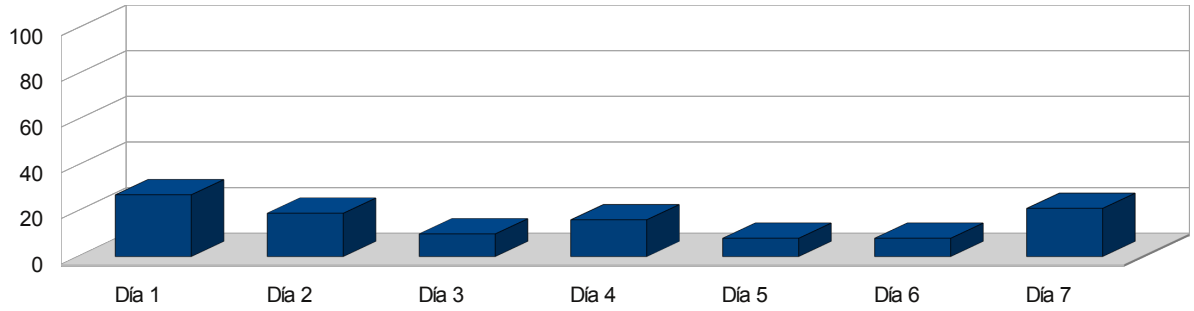

Figura 4.3. Número de noticias de interés por día del usuario Usuario 6

En la figura 4.4, para las métricas de precisión, éstas se mantienen constantes a lo largo del tiempo, aunque se obtiene el máximo valor con poco más del 60%. La métrica de MAP tiende a disminuir. Estos resultados se pueden justificar debido a la falta de información en el propio perfil, ya que no se tiene suficiente realimentación.

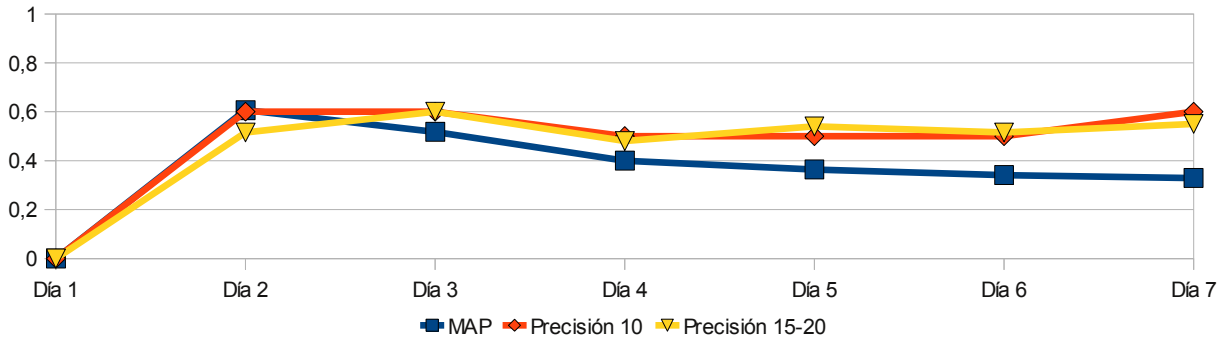

Figura 4.4. Progreso durante una semana del usuario Usuario 6

4.3 Usuario Usuario 10

Como muestra la figura 4.5, en toda la semana el interés no es mayor del 20%, como consecuencia, tampoco se tendrá demasiada información para definir su perfil.

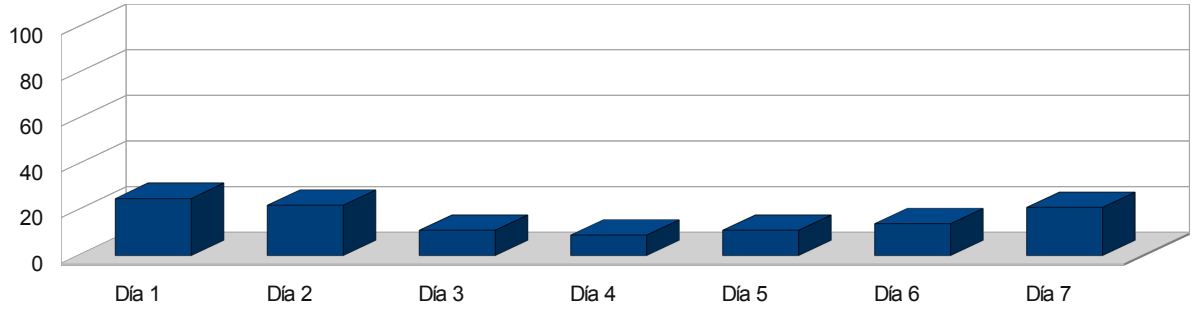

Figura 4.5. Número de noticias de interés por día del usuario Usuario 10

A pesar de tener tan poca información la métrica de precisión sobre las 10 primeras noticias se mantiene sobre el 70%, llegando a valer un poco más del 80%. La precisión sobre entre las 15 y 20 primeras noticias va aumentando de forma constante cada día para casi converger con la P10. El MAP tiene una importe disminución con respecto al primer día, pero, logra mantenerse en un valor constante por encima del 40%.

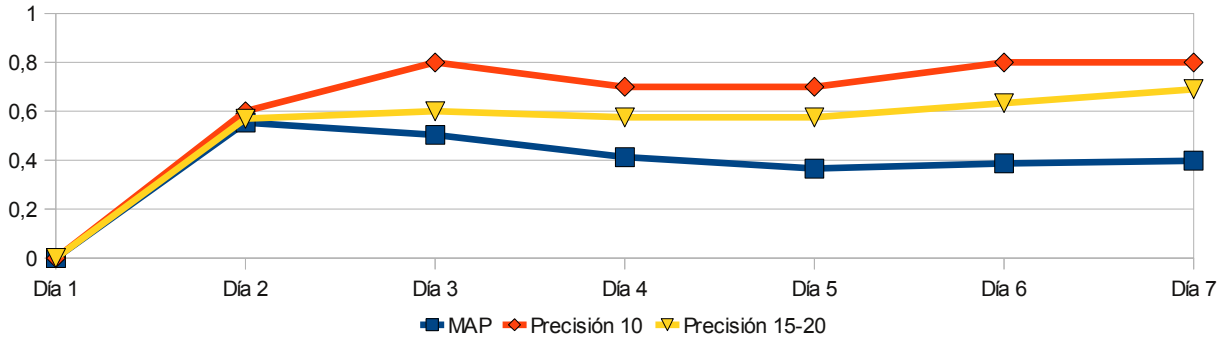

Figura 4.6. Progreso durante una semana del usuario Usuario 10

4.4 Usuario Usuario 8

Este es el usuario con menos interés sobre el conjunto de las noticias, en ningún día alcanza más del 15% de interés sobre todo el conjunto, como muestra la figura 4.7.

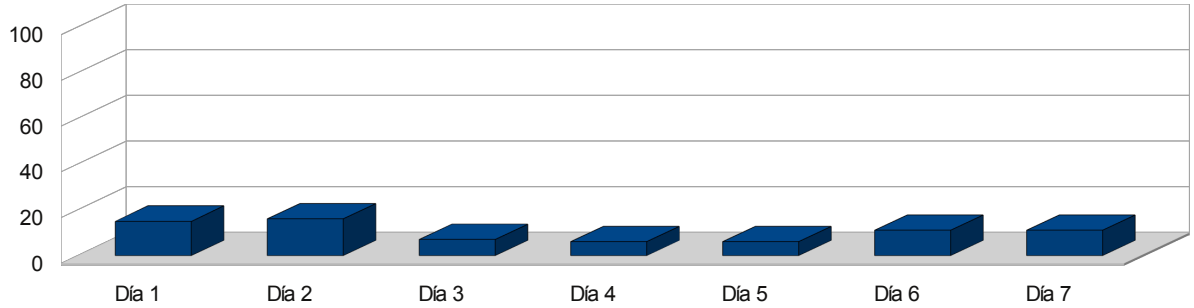

Figura 4.7. Número de noticias de interés por día del usuario Usuario 8

En la figura 4.8 se destaca el pico del segundo día de una precisión del 100% para las 10 primeras noticias y un MAP del 80%. Sin embargo, la tendencia de todas las métricas es a decrecer hasta que las medidas de precisión convergen en el 60%, manteniendo este valor, y la métrica MAP, sólo tiende a disminuir.

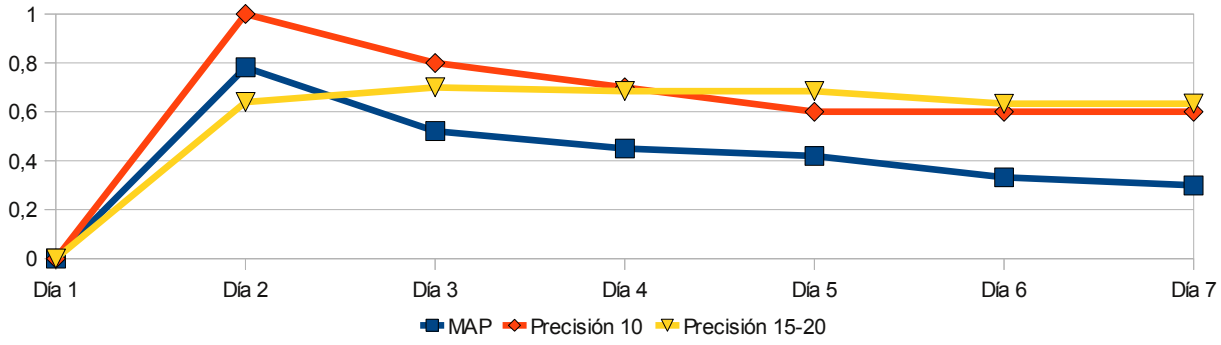

Figura 4.8. Progreso durante una semana del usuario Usuario 8